

Measuring brands in machine answers: A scoping review and validity framework for large language model brand-visibility audits

Charbel Sami Ibrahim

charbel.s.ibrahim@gmail.com

Independent Researcher, Byblos <https://orcid.org/0009-0006-4646-3571>

Systematic Review

Keywords: generative AI, large language models, brand visibility, consideration set, measurement validity, generative engine optimisation

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11390685/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: The author is Social Media Manager at One (withone.ai), a company that builds infrastructure for AI agents; operates Cloud, a social media marketing agency; and provides digital marketing services to Hanks Belts & Leather Goods, a consumer brand. These roles are in AI-related and marketing-services businesses that could be perceived as relevant to the topic of this review. None of these organisations funded the study, which received no funding.

Abstract

Firms increasingly track how often large language models (LLMs) and AI answer engines mention, rank and recommend their brands, yet the audits behind these “AI visibility” metrics reach conflicting conclusions. This scoping review, reported under PRISMA-ScR, maps 38 empirical studies (2024–2026) that query LLMs and measure outcomes at the brand or business level, and appraises them against nine measurement-validity criteria drawn from marketing measurement theory. The evidence base is young and fragile: 35 studies are not peer reviewed, half come from authors with commercial stakes in visibility services, and two vendor groups account for 12 papers. Few studies validate scores against external brand or market benchmarks (2) or repeat measurement over time (7), and results are sensitive to paraphrasing, query language and the number of draws. Drawing on consideration-set theory, I propose a two-stage account that can reconcile many apparent contradictions: inclusion in an answer (membership) appears to track how extensively a brand is documented online, while selection among included brands may be more responsive to attribute evidence, a conjecture the evidence supports only partly. Designs that supply a candidate list observe only selection. Six propositions and a minimum design and reporting standard for brand-visibility audits follow.

Introduction

Conversational AI systems are becoming a route to product discovery. When a consumer asks ChatGPT, Gemini, Claude or Perplexity for “the best CRM for a small agency” or “a good waterproof hiking jacket”, the system answers with a short set of named brands, often ranked and justified. For marketers this answer is a new and largely opaque touchpoint that can shape which brands a buyer considers before any website is visited (Sarstedt et al 2026), and observational evidence suggests that such recommendations are followed by shifts in subsequent browsing (Iannelli and Ai 2026). A market for “generative engine optimisation” (GEO) and “AI visibility” tools has formed around the promise of measuring and improving a brand’s presence in these answers (Aggarwal et al 2024), and audits of how LLMs recommend brands have multiplied since 2024.

These audits do not agree. Some conclude that LLMs favour incumbents: well-known brands are named in most answers while challengers are absent (Kumar 2026; Jack et al 2026d). Others find that brand prominence in answers is only weakly related to established measures of brand strength (Malthouse et al 2026), that a marginal quality advantage is enough to displace a famous brand (Chu and Hou 2026), or that what a model says about a brand depends more on the language of the question than on the brand (Žatuchin 2026g). Still others report that the same prompt repeated later the same day returns a substantially different set of brands (Schulte et al 2026). For a manager reading a vendor dashboard, and for a researcher designing the next study, it is not obvious whether these are substantive disagreements about how LLMs treat brands or artefacts of how each audit was built.

The question matters because audit numbers are already marketed as brand metrics, under labels such as “share of model”, “AI visibility score” and “recommendation share”. Marketing has well-developed

standards for what makes a metric valid (Churchill 1979; MacKenzie et al 2011) and a long tradition of measuring the constructs these audits target: brand salience and mental availability (Romaniuk and Sharp 2004) and consideration-set membership (Hauser and Wernerfelt 1990; Nedungadi 1990; Shocker et al 1991). These standards have not yet been applied systematically to the new evidence base. A critical survey of GEO research focuses on whether content is retrieved and cited by generative engines rather than on brand-level outcomes (Martinez 2026a). A companion methodological paper shows how the choice and weighting of prompts defines the “answer market” that any visibility score summarises, but reports no audit of existing studies (Martinez 2026b). Several proposals specify brand-level scores, including a generative visibility score grounded in mental availability (Lüders 2026), an open specification separating mention visibility, factual alignment and sentiment (Khan 2026), and a per-entity profile distinguishing raw from verified mentions (Varga 2026); these define measures but do not appraise the designs of the studies that produce the evidence. Table 1 positions the present review against this work.

Table 1
Positioning of this review against related syntheses and measurement proposals

Work	Type	Unit of analysis	What it adds	What this review adds
Martinez (2026a)	Critical survey of GEO (45 studies)	Content and citations	Map of optimisation methods and their evidence	Brand-level outcomes; appraisal of audit designs
Martinez (2026b)	Methodological framework, no new data	GEO scores incl. brand mentions	Prompt corpora and weights define the measured market; partial identification	Applies such concerns to an evidence base of 38 brand audits and links them to marketing constructs
Lüders (2026)	Metric proposal with data	Brand salience in LLMs	Parametric versus inference-time salience	Cross-study validity appraisal; validation status of such scores
Khan (2026)	Open measurement specification	Brand mentions	Separable visibility, alignment and sentiment measures	Evidence on which designs produce reliable scores
Žatuchin (2026d); Schulte et al (2026)	Reliability protocols	Repeated queries	How many runs a stable score requires	Reliability as one of nine criteria; consideration-set account of conflicting findings
This review	Scoping review and validity appraisal	Brand and business outcomes	Evidence map, nine-criterion appraisal, two-stage propositions, reporting standard	—

This article addresses that gap with a scoping review of empirical studies that query LLMs or AI answer engines and record outcomes at the level of brands, companies, businesses or branded products. Three research questions guide the review:

- RQ1. How do empirical studies operationalise a brand’s “visibility” in the outputs of LLMs and AI answer engines?
- RQ2. Which design choices vary across studies, and are divergent substantive conclusions associated with them?
- RQ3. Judged against marketing measurement standards, which validity threats recur, and what minimum design and reporting standards follow?

The review makes three contributions. First, to my knowledge it provides the first coded evidence map and validity appraisal of brand-level LLM audits, covering 38 studies, together with an open evidence register. Second, it offers a two-stage account grounded in consideration-set theory: brand *membership* in an answer and brand *selection* within it are distinct outcomes that may have different drivers, and audit paradigms differ in which stage they can observe. Six propositions state the account in testable form, with the strength of evidence for each made explicit. Third, it translates nine validity criteria into a minimum design and reporting standard that researchers, reviewers and managers can apply to any brand-visibility audit, including commercial dashboards.

Conceptual background

LLM answers as a brand channel

General-purpose LLMs produce brand-relevant output in two ways. In parametric mode the model answers from associations learned during training, without consulting external sources. In retrieval-augmented mode the system issues its own web searches, reads selected pages and composes an answer that may cite them; consumer products such as ChatGPT with search, Perplexity, Gemini and Google’s AI Overviews and AI Mode increasingly work this way. The distinction matters for marketers because the levers differ. A brand’s presence in parametric answers reflects the historical volume and character of text about it, a slow-moving stock, whereas its presence in retrieval-augmented answers depends on what the system’s searches return at the time of the query, a flow that content, public relations and technical accessibility can influence (Chen and Koudas 2026; Hamann et al 2026).

Both modes interpose a new agent between the consumer and the market. The answer is a short list rather than a page of links, and it arrives with reasons attached. In the terms of choice architecture, the system preselects and frames options before the human decision-maker sees them (Sarstedt et al 2026). Laboratory and field evidence suggests that such recommendations are associated with changes in human spending and browsing (Zac and Gal 2025; Iannelli and Ai 2026); this review leaves those downstream effects aside and focuses on how brand presence in answers is measured.

From mental availability to machine availability

Two marketing constructs are the natural reference points for audit outcomes. The first is the consideration set: the subset of brands a consumer actively evaluates before choice. Consideration-set research established that choice can be described as a two-stage process in which brands must first enter the set, through recall or exposure, and are then compared on attributes (Hauser and Wernerfelt 1990; Shocker et al 1991). Entry and comparison follow different logics: Nedungadi (1990) showed that making a brand easier to recall can change choice without changing its evaluation. An LLM answer listing five brands performs a function similar to a consideration set. The analogy does not imply that a model reasons in two steps, since inclusion and ranking emerge from a single generation; it implies that inclusion and ranking are distinct observable outcomes that should be measured separately rather than combined into one score.

The second construct is brand salience, or mental availability: the propensity of a brand to come to mind in buying situations (Romaniuk and Sharp 2004). Salience is distinct from brand attitude and is measured across many category entry points rather than a single cue. If LLM answers inherit the distribution of text about brands, regularities familiar from brand-performance research, such as the double-jeopardy pattern in which smaller brands score lower on several measures at once (Ehrenberg et al 1990), may appear in machine answers too, and recommender-systems research gives reason to expect that algorithmic intermediaries can amplify popularity differences (Fleder and Hosanagar 2009). Whether they do, and under which designs, is the empirical question the reviewed audits address. Customer-based brand equity frameworks add a further distinction: awareness (recognition and recall) is necessary but distinct from the strength, favourability and uniqueness of associations (Keller 1993). An audit that counts mentions measures something closer to recall; an audit that codes what is said measures associations.

Audit-based brand metrics as measurement

Brand-visibility audits are algorithm audits: they infer the behaviour of a system that cannot be inspected directly by sending it controlled inputs and recording outputs (Metaxa et al 2021). Audit designs inherit known hazards. Outputs vary with location and personalisation (Hannak et al 2013), the population of queries sent determines what can be generalised, and the system itself changes over time. LLMs add a further hazard because their outputs are sampled: the same input can produce different answers, so any single response is one draw from a distribution.

Marketing measurement theory supplies the yardstick for judging such metrics. A valid measure requires a clearly defined construct, a sampling of the domain that represents it, evidence of reliability, and evidence that scores relate to other variables as theory predicts (Churchill 1979; Messick 1995; MacKenzie et al 2011). Applied to audits, these requirements become concrete design questions: what exactly is measured (inclusion, rank or favourability); whether the prompts represent the questions real buyers ask; whether enough draws are taken to estimate a stable rate; whether the brand universe is

defined independently of the model's own outputs; whether brand names are identified accurately in free text; whether scores relate to external brand metrics; and whether they are stable over time. Generalisability theory offers a vocabulary for the reliability questions, decomposing score variance into facets such as prompts, models, languages and occasions; several reviewed studies use it explicitly (Žatuchin 2026d, 2026g). These requirements form the appraisal criteria described next.

Method

Design and protocol

I conducted a scoping review following Arksey and O'Malley (2005) and report it according to the PRISMA extension for scoping reviews (PRISMA-ScR; Tricco et al 2018); the completed checklist is in Online Appendix A. A scoping design suits a field in which designs and outcome definitions are heterogeneous and most evidence is recent, so the aim is to map and appraise rather than to pool effect sizes. The protocol, covering research questions, eligibility criteria, search strategy, charting fields and appraisal criteria, was written on 15 October 2026 immediately before screening began, on the same day as the searches. It was not registered. Eleven deviations from it are logged with reasons in Online Appendix A; those that affect interpretation are reported below.

Eligibility criteria

Studies were eligible if they (a) systematically queried general-purpose LLMs, LLM-based assistants or AI answer engines; (b) recorded at least one outcome at the level of brands, companies, businesses or branded products, such as mention, inclusion, rank, share, sentiment or attributes of named brands, or citation of brand-owned sources; (c) reported original empirical data; (d) were posted or published between 1 January 2023 and the search date; and (e) were written in English. Peer-reviewed articles, conference papers, preprints, working papers, theses and technical reports with a described method were all eligible, because most of this literature circulates outside journals; publication status and commercial interests were charted and are taken into account in the synthesis.

During screening the concept was broadened to include individual businesses, such as restaurants and local service firms, which compete as brands in their markets, and an exclusion was added for studies using only fictitious or hypothetical brands; both changes are logged. Excluded were content-level GEO studies whose outcome is the retrieval or citation of documents without a brand-level outcome; studies using only fictitious or hypothetical brands; experiments on human responses to LLM recommendations (retained as context); conceptual papers and measurement proposals without audit data (retained for theory); vendor reports without a described method; and studies using LLMs as research instruments, for example as synthetic respondents or analysts, rather than auditing their outputs as a market channel.

The protocol required openly accessible full text. In practice, much of this literature is posted as working papers or vendor reports whose full text sits behind automated bot checks. Studies whose full text could not be read but whose abstract or landing page reported design details sufficient for charting (systems,

sample, outcome and principal result) were therefore included as a separate, flagged tier (abstract-level, $n = 13$). Records for which neither the full text nor such an abstract could be reached without passing an interactive challenge were counted as not retrievable. The same rule was applied to all records.

Information sources and search

Subscription databases (Scopus, Web of Science, Business Source, ABI/INFORM) were not available, so the search relied on open sources where this literature concentrates. Seven initial searches of arXiv (S01–S07) combined system terms (“LLM”, “ChatGPT”, “AI search”, “generative engine”) with brand-outcome terms (“brand”, “brand bias”, “product recommendations”, “AI visibility”), screening the first page of results sorted by date (85 results). Seven searches of Google Scholar (S08–S14), limited to 2023 onwards, added terms from the marketing literature (“consideration set”, “share of voice”, “share of model”, “brand visibility”, “AI Overviews”) and captured SSRN working papers, conference proceedings and journal articles; the first 10–20 results per string were screened (80 results). After an internal review of the first draft, three supplementary arXiv abstract-field searches were run and their complete result lists screened (“brand LLM”, 126 results; “brands ‘language models’ recommendations”, 25; “generative engine optimization”, 53). Citation searching combined backward chasing of included studies with forward searches of two widely cited studies through the Semantic Scholar database (24 citing records). Further forward searches could not be completed because of rate limits and an interactive check that was not circumvented. All searches were run on 15 October 2026; strings, sources and counts are in Online Appendix A.

Selection and charting

Screening and charting were carried out by a general-purpose LLM acting as the single screener under the author’s direction (see the section on the use of generative AI below). Every record that passed the title stage was logged with a decision and reason. Title-stage exclusions were counted but not logged individually, and the stage at which content-based exclusions were made was not recorded separately. To check the reliability of this process, a second, independent instance of the LLM, given only the eligibility criteria and record identifiers, re-screened a random 20% sample of the screened records (18 of 89) without access to the first decisions. The two agreed on 16 of 18 records (Cohen’s kappa = 0.81 across include, exclude, context and not-retrievable decisions); both disagreements arose because the second screener could not open a page that the first had read, and on the 13 records both could read, agreement was complete. The second instance also independently scored the nine validity criteria for five randomly drawn included studies: scores agreed exactly in 56% of cells and within one scale point in 89%, so the appraisal is best read as a coarse profile. No human second reviewer was involved.

A charting form recorded bibliographic details, publication status, declared or evident commercial interest, systems and versions, access mode (programming interface or consumer product) and retrieval setting, dates of data collection, markets and languages, categories and brands, prompt types and

numbers, repetitions per prompt, sampling temperature, definition of the competitive set, metrics, external benchmarks, temporal design, data availability, principal brand-level findings and stated limitations. All reading was done by the LLM. Of the 25 full texts, 12 were read in the browser as continuous text, five of them only through the methods and main results (approximately the first 50,000 characters), and 13 were charted from a structured extraction of targeted sections of the arXiv full text. Figures quoted in this article were checked against the source abstract where the abstract reports them and otherwise against a second reading of the relevant section; checks are logged in Online Appendix B.

Each study was coded on four derived variables. *Paradigm*: open (the model generates brand names in response to category or need-based prompts), supplied (the researcher places a candidate list in the prompt), named (the focal brand is named in the query), or reanalysis of earlier audit data. *Mode*: parametric, retrieval-augmented, mixed, simulated retrieval through injected documents, or not reported. *Metric family*. *Conclusion family*, coded inductively: prominence advantage in inclusion (A1); brand identity shifting selection or rank (A2); prominence advantage attenuated, that is, weaker than market position would predict or conditional on attribute parity (A3); dependence on retrieval or off-site documentation (B); instability or design sensitivity of the metric (C); gating by language, market or model origin (D); manipulability through content or injection (E); and divergence between raw and verified or endorsed mentions (F). A study could carry several codes.

Validity appraisal

Each included study was appraised on nine criteria derived from the measurement literature (Table 2). Criteria V1–V8 were fixed in the protocol; V9, the validity of brand-name extraction, was added during charting after several studies reported that alias matching or language-model judges had misidentified brands, in one case inflating a firm’s count about sevenfold before correction (Žatuchin 2026h). Each criterion was scored Yes, Partly, No, Not reported (NR) or Not applicable (NA); V9 is NA for supplied-list and forced-choice designs, in which identifying the chosen brand is trivial. Decision rules were tightened after the internal review (for example, V2 = Yes requires at least five draws per prompt and condition for the main outcome; V6 = Yes requires an external brand or market benchmark, with proxies scored Partly) and all studies were re-scored. For description only, a summary score expresses points earned (Yes = 2, Partly = 1, otherwise 0) as a percentage of the applicable maximum; because NR is scored like No, the summary score penalises incomplete reporting as well as weak design, and profiles across criteria are more informative than the summary.

Table 2
Validity criteria applied to brand-visibility audits

Criterion	Question asked of each study	Grounding
V1 Construct definition	Is “visibility” defined, and are inclusion, prominence (rank) and favourability kept distinct?	Churchill (1979); Keller (1993); Romaniuk and Sharp (2004)
V2 Stochastic sampling	Are at least five draws taken per prompt and condition, and is variance reported?	Reliability; generalisability theory
V3 Prompt representativeness	Is the prompt source justified (for example, buyer queries) and are phrasings varied?	Content validity; domain sampling (Churchill 1979)
V4 Independent brand set	Is the brand universe defined outside the model’s outputs?	Consideration-set measurement (Shocker et al 1991)
V5 System specification	Are model and version, access mode, retrieval setting, location and dates reported?	Audit reproducibility (Metaxa et al 2021)
V6 Criterion validity	Are scores related to an external brand or market benchmark (brand equity, market share or revenue)?	Criterion and nomological validity (MacKenzie et al 2011)
V7 Temporal stability	Is measurement repeated over time and drift separated from noise?	Test–retest reliability
V8 Transparency	Are prompts, raw outputs and code available?	Open science; audit accountability
V9 Entity-extraction validity	Is the identification of brand names in free text validated against human coding (precision, recall)?	Measurement error in coding (added during charting)

Use of generative AI in conducting the review

The author used a large language model (Claude, Anthropic; October 2026) to assist with the review design, literature searching, screening, data charting, analysis, figures and drafting. Reliability was checked with a blind second instance of the model, as described above. The author reviewed, revised and approved all content; the interpretations reflect his own views, and he takes full responsibility for the manuscript.

Synthesis

The synthesis comprises a descriptive map of the evidence (counts by design feature), cross-tabulations of conclusion families by paradigm and by commercial interest (Table 4), a narrative synthesis organised by research question, and a sensitivity check that repeats the main counts without the abstract-level studies. Effect sizes were not pooled because metrics, units and denominators differ. Because two research groups contributed eight and four papers respectively, some of which reuse data, patterns are described in terms of the number of independent groups supporting them as well as the number of

papers. The charting data, screening log and search log are provided as an open evidence register (Online Appendix B).

Results

Search results and profile of the evidence

The searches identified 402 records: 289 from arXiv, 80 from Google Scholar and 33 from citation searching (Fig. 1). After removal of 112 duplicates, 290 titles were screened and 201 excluded as clearly irrelevant. Of 89 abstracts screened, 39 were excluded (23 had no brand-level outcome or were otherwise out of scope, 15 were conceptual papers or measurement proposals without audit data, and 1 was a vendor document without a method). Of 50 reports sought, 9 could not be retrieved. Of 41 assessed, 2 were retained as context because they study human responses to LLM recommendations and 1 was a late-identified duplicate. Thirty-eight studies were included: 25 charted from full text and 13 at abstract level. Table 3 summarises them; the full charting is in the evidence register.

The evidence base is recent, mostly unrefereed and partly commercial. Thirty-two of the 38 studies date from 2026, four from 2025 and two from 2024. Only three appeared in peer-reviewed proceedings, all at one computational linguistics venue (Kamruzzaman et al 2024; Pfrommer et al 2024; Filandrianos et al 2025); one is described by its author as a non-archival workshop paper and the remainder are preprints, working papers or technical reports. Nineteen studies were written by authors who sell GEO or AI-visibility services or whose own employer's brand is among those measured. The evidence is also clustered: one author affiliated with a visibility vendor contributed eight papers, one of which reanalyses data from the others, and one vendor team contributed four papers that partly share data, so the 38 papers come from 27 research groups. Several sources carry correction notices: one withdrew results that had partly come from a synthetic-data generator (Žatuchin 2026d), and another corrected a brand-matching error that had inflated one firm's count about sevenfold (Žatuchin 2026h). Citation errors also occur in sources, such as an author list attached to the wrong arXiv identifier and an incorrect page range for a peer-reviewed paper. None of these features is disqualifying in a fast-moving field, but together they mean that headline numbers from single studies should be read with caution, and that patterns replicated across independent groups deserve more weight than any single estimate.

Designs cluster in a few configurations. Twenty-seven studies used open generation, in which the model names brands in response to category or need-based prompts (one combined this with named-brand queries and one with fixed venue lists); six supplied a candidate list; four named the focal brand in the query; and one reanalysed earlier audit data. Fifteen studies audited retrieval-augmented systems, nine mixed parametric and retrieval settings, seven were purely parametric, two simulated retrieval by placing documents in the prompt, and five did not report the setting. Where access was reported, most studies queried programming interfaces; only three observed the consumer products that buyers use directly (Schulte et al 2026; Žatuchin 2026f; Budiman 2026). Thirteen used a language other than English.

Sixteen studies took at least five draws per prompt and condition and reported variance, nine took fewer or did so only for a subset, four relied on a single response, and nine did not report the number.

Table 3
Included studies (N = 38)

Study	Status	Design / mode	Draws	Principal brand-level finding	Score (%)
Arshad (2026)	PP, COI, AB	open / retrieval	3	Same brands cited 16% to 91% depending on platform	44
Budiman (2026)	PP, AB	open / retrieval	1 ×32 users	Stable core of national brands; disagreement in the tail	39
Chen et al (2025)	PP	open / retrieval	1?	Earned media dominate citations; popularity-framed prompts favour majors	22
Chen et al (2026)	PP, AB	open+supplied / NR	NR	47.5% of real restaurants never recommended even from a fixed list; fabrication clusters in weakly documented neighbourhoods	28
Chen and Koudas (2026)	PP	open / retrieval	daily	Retrieval explains citations; big brands enter via model prior, niche brands via third-party pages	83
Chu and Hou (2026)	PP	supplied / parametric	20–30	Incumbent chosen only when options are identical; a small rating edge overturns it	69
Cook (2025)	PP, COI, AB	open / NR	NR	Visible, English-content-rich firms over-represented	22
Filandrianos et al (2025)	PR	supplied / parametric	100	Persuasion cues in listings shift recommendation rates	50
Gonzalez Castro (2026)	PP, AB	open / NR	8	Raw presence 38%, but most mentions flag the brand's changed status	67
Hamann et al (2026)	PP, AB	named / retrieval	NR	Brand knowledge in LLM answers updates through retrieval	44
Notes: PR = peer reviewed; PP = preprint, working paper or technical report; COI = authors have a commercial interest in AI-visibility services or in a measured brand; AB = charted at abstract level. Design: open = model generates brand names; supplied = candidate list in prompt; named = focal brand named in query. Draws = responses per prompt and condition (* for part of the design only; 1 × 32 users = one run by each of 32 people). Score = validity points earned as a percentage of the applicable maximum (Table 2); NR is scored as 0, so low scores reflect incomplete reporting as well as design. Full details in Online Appendix B.					

Study	Status	Design / mode	Draws	Principal brand-level finding	Score (%)
House (2026)	PP, COI, AB	open / mixed	NR	70% of businesses never mentioned; off-page signals matter in 3 of 5 systems	50
Huang et al (2025)	PP, COI	open / mixed	NR	Chinese-developed models mention brands more often (88.9% vs 58.3% of answers)	28
Jack et al (2026a)	PP, COI	open / retrieval	≥ 3	Providers disagree on sets (Jaccard ≈ 0.35); many picks lack retrieval support	44
Jack et al (2026b)	PP, COI	open / retrieval	20–30	Paraphrasing cuts set overlap to 0.29 (cosmetic) and 0.14 (constraint-adding)	39
Jack et al (2026c)	PP, COI, AB	open / retrieval	10	Personas reshuffle mid-market brands; leaders persona-resistant	33
Jack et al (2026d)	PP, COI	open / retrieval	≤ 30	Inclusion falls from 77% (leaders) to 3% (tail); conversion once included is flatter	56
Kamruzzaman et al (2024)	PR	supplied / parametric	1	Global brands linked to positive and local brands to negative attributes	38
Kireyev (2026)	PP	supplied / parametric	NR	About half of models prefer major hotel chains to otherwise identical independents; weak models follow list order	50
Kumar (2026)	PP, COI	open / mixed	10+	First-run inclusion 73% / 44% / 11% by brand-stature tier; most cells all-or-nothing	56
Kurtskhalia (2026)	PP, AB	open / parametric	multi	Quantised serving raises same-brand repetition in a small model	33
Lüders (2026)	PP, AB	named / mixed	NR	Proposes parametric vs inference salience score;	22
Notes: PR = peer reviewed; PP = preprint, working paper or technical report; COI = authors have a commercial interest in AI-visibility services or in a measured brand; AB = charted at abstract level. Design: open = model generates brand names; supplied = candidate list in prompt; named = focal brand named in query. Draws = responses per prompt and condition (* for part of the design only; 1 × 32 users = one run by each of 32 people). Score = validity points earned as a percentage of the applicable maximum (Table 2); NR is scored as 0, so low scores reflect incomplete reporting as well as design. Full details in Online Appendix B.					

Study	Status	Design / mode	Draws	Principal brand-level finding	Score (%)
validation pending					
Malthouse et al (2026)	PP	open / parametric	40*	Many established brands never listed; prominence tracks search interest more than brand-equity rankings (exploratory)	67
Paeng (2026)	PP	supplied / simulated-retrieval	50–100	Injected manipulation demotes the brand in some models, promotes it in others	62
Pfrommer et al (2024)	PR	supplied / simulated-retrieval	10	Brand names drive rankings; injected text promotes low-ranked products	50
Pitenin (2026)	PP, AB	open / retrieval	NR	86% of 4,776 venues never recommended; entry follows documentation, first position follows rating	67
Rienecker et al (2026)	PP	open / NR	5	Top entity in 70–76% of answers; US models over-recommend US entities	50
Schulte et al (2026)	PP	open / retrieval	≤ 10/day	Brand-set overlap 0.33–0.48 within a day; ≥7 runs per prompt per day needed	50
Sharma (2026)	PP	open+named / mixed	1	Startups known when named (94–99%) but rarely surfaced in discovery (3–8%)	39
Tannenbaum (2026)	PP, COI	open / retrieval	≥ 2	Mention ≈ 3% without own-domain retrieval vs 91–100% with it	44
Varga (2026)	PP, COI	named / parametric	1	Raw mentions overstate verified visibility; fabricated sources common	50
Zatuchin (2026a)	PP, COI, AB	open / NR	NR	Recipient gender and season change number and identity of brands	22

Notes: PR = peer reviewed; PP = preprint, working paper or technical report; COI = authors have a commercial interest in AI-visibility services or in a measured brand; AB = charted at abstract level. Design: open = model generates brand names; supplied = candidate list in prompt; named = focal brand named in query. Draws = responses per prompt and condition (* for part of the design only; 1 × 32 users = one run by each of 32 people). Score = validity points earned as a percentage of the applicable maximum (Table 2); NR is scored as 0, so low scores reflect incomplete reporting as well as design. Full details in Online Appendix B.

Study	Status	Design / mode	Draws	Principal brand-level finding	Score (%)
Žatuchin (2026b)	PP, COI, AB	named / retrieval	NR	86% of citations for brand questions point to third-party sites	11
Žatuchin (2026c)	PP, COI	open / mixed	1–40	Brand-behaviour profiles of systems do not transfer across domains	39
Žatuchin (2026d)	PP, COI	reanalysis / mixed	5–40	Required runs depend on prompt design (G = 0.96 vs 0.27); single cells underpowered	56
Žatuchin (2026e)	PP, COI	open / retrieval	1 (5*)	Home-language queries raise local champions' share far more than global brands'	50
Žatuchin (2026f)	PP, COI	open / retrieval	6	Query language gates local suppliers; IP selects market	44
Žatuchin (2026g)	PP, COI	open / mixed	1 (≈ 5*)	Language explains 26–32% of sentiment variance; single answers unreliable	39
Žatuchin (2026h)	PP, COI	open / mixed	5	Inclusion share tracks firm size but is less concentrated than revenue among sampled brands	89
<i>Notes:</i> PR = peer reviewed; PP = preprint, working paper or technical report; COI = authors have a commercial interest in AI-visibility services or in a measured brand; AB = charted at abstract level. Design: open = model generates brand names; supplied = candidate list in prompt; named = focal brand named in query. Draws = responses per prompt and condition (* for part of the design only; 1 × 32 users = one run by each of 32 people). Score = validity points earned as a percentage of the applicable maximum (Table 2); NR is scored as 0, so low scores reflect incomplete reporting as well as design. Full details in Online Appendix B.					

RQ1: How visibility is operationalised

The studies measure at least five different things under the label of visibility, and they rarely say which one they mean.

Inclusion. The most common outcome, used in 16 studies, is whether a brand or business appears in an answer. It goes by many names, including mention rate, AI presence, category inclusion rate, preferred-entity inclusion rate, surface rate and brand retrieval prevalence (Malthouse et al 2026; Rienecker et al 2026; Gonzalez Castro 2026; Žatuchin 2026h; Jack et al 2026d). Identical labels hide different denominators: the share of all answers, the share of runs for a given prompt, or the share of prompts in which a brand appears at least once. These are not interchangeable. In one large commercial dataset, 63% of brand–prompt–engine cells never mentioned the brand and 14% always did (Kumar 2026), so an

average “mention rate” of 30% may describe a brand that is reliably present for a third of buyer questions and absent from the rest, not one that appears in a third of answers to every question.

Prominence. Six studies measure position or rank within an answer, through mean reciprocal rank (Malthouse et al 2026), average position (Kumar 2026), first position (Pitenin 2026) or rank scores in supplied lists (Pfrommer et al 2024).

Selection and share. Studies using supplied lists measure which option is chosen (Chu and Hou 2026; Filandrianos et al 2025; Paeng 2026; Kireyev 2026). Open-generation studies compute shares across brands and their concentration (Žatuchin 2026h; Kurtskhalia 2026), or separate being mentioned from being recommended (Jack et al 2026d).

Representation. Eight studies code what is said, including sentiment, attribute associations, conveyed brand knowledge, accuracy against the brand’s own facts, or whether a mention is endorsed, hedged or flagged as outdated (Kamruzzaman et al 2024; Varga 2026; Gonzalez Castro 2026; Hamann et al 2026; Žatuchin 2026g).

Sources. Seven studies record which web pages are cited when a brand is discussed, and classify them as brand-owned, earned or social (Chen et al 2025; Žatuchin 2026b; Budiman 2026; Arshad 2026).

A further eight studies measure the *stability* of these outcomes across repeated runs, paraphrases, personas or providers, which is a property of the measurement rather than of the brand. Proposed composite scores add to the variety (Lüders 2026; Varga 2026; Khan 2026). Only 17 studies defined their construct clearly enough to keep inclusion, prominence and favourability apart (V1).

The variety is not a defect in itself; inclusion and favourability are different constructs and both matter. The difficulty is that results are compared, aggregated and sold as though they measured the same thing. Two observations show why this matters. Raw mention counts can misstate visibility in a substantive sense: when every mention of a revived retail brand was read, 88% carried a disclaimer about the brand’s changed status and only 1.7% treated the old entity as current (Gonzalez Castro 2026), and answers about high-salience firms were accompanied by fabricated sources more often than answers about low-salience firms (Varga 2026). And the same brands were cited in 16% of answers on one platform and 91% on another (Arshad 2026), so a platform-pooled score is partly a statement about the platform mix.

RQ2: Design choices and divergent conclusions

Table 4 cross-tabulates conclusion families by paradigm and by commercial interest. Three patterns stand out. Prominence advantage in inclusion (A1) is reported only by open-generation studies, whereas brand identity shifting selection (A2) and manipulability (E) are reported mostly by supplied-list studies. Instability (C) and language or market gating (D) are reported predominantly by authors with commercial interests (11 of 14 and 7 of 10 studies), a pattern consistent with the services those authors sell, and one that calls for independent replication rather than dismissal. All five manipulability findings and all

five A2 findings come from studies without commercial interests. The sections below examine how far design explains the apparent contradictions.

Table 4
Conclusion families by design paradigm and commercial interest (number of studies)

Conclusion family	All (38)	Full text (25)	Open (27)	Supplied (6)	Named or reanalysis (5)	COI (19)	No COI (19)
A1 Prominence advantage in inclusion	11	7	11	0	0	6	5
A2 Brand identity shifts selection or rank	5	5	1	4	0	0	5
A3 Advantage attenuated or conditional	6	4	5	1	0	2	4
B Retrieval or documentation dependence	11	6	9	0	2	4	7
C Instability or design sensitivity	14	8	13	0	1	11	3
D Language, market or origin gating	10	6	8	1	1	7	3
E Manipulability	5	5	0	5	0	0	5
F Raw versus verified mentions	2	1	1	0	1	1	1
Note: A study can carry several codes. COI = commercial interest in AI-visibility services or a measured brand.							

Supplied versus open designs

The clearest apparent contradiction concerns incumbent advantage. In open-generation designs, prominent brands dominate inclusion. Brand-stature tiers had first-run inclusion rates of 73%, 44% and 11% in a large commercial dataset (Kumar 2026). In an audit of about 37,000 runs with an external catalogue of 533 brands, the probability of surfacing in a given query fell from 77% for category leaders to 3% for the least prominent tier, and about half of the two lowest tiers never surfaced (Jack et al 2026d). Startups listed on Product Hunt were recognised 94–99% of the time when named, yet surfaced in only 3–8% of discovery-style queries (Sharma 2026). Studies of businesses report the same steep exclusion: 69.5% of 2,729 small and local businesses were never mentioned by any of five systems (House 2026), and in a complete census of 4,776 restaurants, cafes and bars, 85.6% were never recommended (Pitenin 2026).

Supplied-list designs reach a more qualified conclusion. When a real incumbent was placed among fictional competitors with identical specifications, models chose it every time, but a rating advantage of less than a tenth of a star ended its dominance, and product attributes explained 82.4% of rank variance against 1.2% for brand identity (Chu and Hou 2026). Other supplied-list studies show that brand names nonetheless shape rankings (Pfrommer et al 2024), that booking agents built on about half of the models tested preferred major hotel chains to otherwise identical independent hotels (Kireyev 2026), that listing order and persuasive copy move choices (Filandrianos et al 2025; Kireyev 2026), and that injected text can promote or demote a brand depending on the model family (Paeng 2026).

Part of the disagreement follows from what each design can observe. A supplied list removes the inclusion problem by construction: every brand is already in the set, so the design measures only selection among included options. Open designs measure the product of inclusion and selection. Two open-generation studies that separate the stages are consistent with different drivers at each. In the restaurant census, entry into answers was associated with how extensively a venue is documented (review volume, own website, price information and third-party mentions), not with its star rating, whereas among recommended venues the rating predicted first position (Pitenin 2026). In the vendor audit, conversion from being mentioned to being recommended was 25–41% for leaders and highest, 37–52%, for challengers (Jack et al 2026d). The second comparison needs care: brands that are included despite low prominence may be unusually well matched to the query, so higher conditional conversion can arise from selection on inclusion rather than from a weaker role of prominence at the selection stage. Evidence on selection is also mixed: revealing brand names reordered otherwise identical products (Chen and Koudas 2026), models relied on product names when ranking supplied products (Pfrommer et al 2024), and chain affiliation carried a choice premium for many models (Kireyev 2026). The defensible conclusion is that inclusion is steeply graded by prominence, while the weight of prominence at selection is smaller in some designs, conditional on attribute parity in others, and not yet well estimated.

What kind of prominence?

The studies operationalise prominence differently: hand-coded stature based on encyclopaedia coverage, press and funding (Kumar 2026); awareness tiers from analyst lists and directories (Jack et al 2026d); revenue (Žatuchin 2026h); brand-equity rankings and search interest (Malthouse et al 2026); and documentation such as reviews, websites and third-party mentions (Pitenin 2026; House 2026). Read together, they suggest that inclusion is associated more closely with how extensively a brand is documented and discussed online than with its market position. Documentation measures such as review volume and search interest also reflect consumer demand, so the contrast drawn here is with market position (revenue or share), not with popularity. In the most detailed parametric study, prominence in answers was most strongly associated with search interest and online conversation and only weakly with brand-equity rankings, and many established brands were never listed under category-only prompts, although that analysis was exploratory and based on 82 brands (Malthouse et al 2026). Among 47 sampled firms with size data, inclusion was correlated with a size benchmark built from

revenue and related figures (pooled rank correlation 0.635), and inclusion shares were less concentrated than the benchmark in every industry studied (Žatuchin 2026h), although a bounded share of mentions is mechanically likely to be less concentrated than revenue, so this comparison is suggestive only. For restaurants, fabricated venues clustered in neighbourhoods with weaker digital and physical footprints, many real venues were never recommended, and a demand simulation suggested that reliance on LLM recommendations would shift visits away from chains towards independent venues (Chen et al 2026). Market leaders are usually also the most documented brands, which explains why prominence and documentation are easily confused; the cases where they diverge are informative for theory.

Parametric versus retrieval-augmented systems

A related divergence concerns what drives visibility. In retrieval-augmented systems, visibility depends on what the system's searches return. In a 90-day panel, retrieval explained 32% of within-website citation variation against 2.8% for on-page features (Chen and Koudas 2026). In a vendor panel, target brands were mentioned in about 3% of answers when neither the brand nor its own domain appeared in the retrieval path, against 91–100% when both did (Tannenbaum 2026). Off-site presence on community, directory and reference sites was associated with mention in three of five systems but not in the other two (House 2026), and for brand questions 85.7% of retrieved citations pointed to third-party rather than brand-owned sites (Žatuchin 2026b).

These findings point to two routes into answers rather than rival claims. Within the same systems, large brands were written into the models' own search queries from prior knowledge far more often than niche vendors, and 46% of brand mentions in one system were not supported by any cited page; niche vendors appeared mainly when the search found third-party pages naming them (Chen and Koudas 2026). Across providers, the share of recommendations without retrieval support ranged from 8–29% for one provider's models to 43–52% for another's (Jack et al 2026a), and brand knowledge conveyed by LLMs after a repositioning updated mainly through retrieval (Hamann et al 2026). These are observational patterns; no study has yet manipulated retrievable evidence for matched brands of differing prominence.

Query language, market and model origin

A third set of differences arises from where and in which language a question is asked. Asking in a brand's home language rather than English raised the recommendation share of local champions by 0.80 against 0.15 for global brands on a 0–1 share scale (Žatuchin 2026e). In a consumer-interface experiment, the language of the question determined whether local suppliers appeared at all, while the user's apparent location selected which national market was assumed (Žatuchin 2026f). In a variance decomposition, query language explained more variance in brand sentiment than model or brand (Žatuchin 2026g). Chinese-developed models mentioned brands in 88.9% of answers against 58.3% for international models, and the authors' employer's brand appeared only in the former (Huang et al 2025). US-developed models over-recommended US entities relative to market benchmarks (Rienecker et al 2026), and associations favoured global over local brands (Kamruzzaman et al 2024). These findings come from four groups, but three of the six studies come from one vendor-affiliated author. Taken

together, they indicate that an English-language audit run from one location measures one market's answer set rather than a brand's visibility in general.

Prompts, draws and providers

The fourth source of divergence is the measurement itself. Same-day reruns of identical prompts produced brand-set overlaps (Jaccard) of 0.33–0.48, most of the instability being sampling noise rather than drift (Schulte et al 2026); a restaurant audit likewise found two-week similarity comparable to same-day reruns (Pitenin 2026). Cosmetic paraphrases cut overlap to 0.29 and constraint-adding paraphrases to 0.14, against a rerun baseline of 0.50–0.61 (Jack et al 2026b). Providers agreed on recommendation sets at a Jaccard of about 0.35 (Jack et al 2026a). Reliability analyses show why single-run audits mislead: a single answer carried almost no brand-discriminating signal in one decomposition (Žatuchin 2026g), and the number of runs needed for a dependable score varied from one or two to more than any reported design used, depending on whether prompts were diverse questions or paraphrases of one question (Žatuchin 2026d). Persona framing and recipient gender changed which brands were named (Žatuchin 2026a; Jack et al 2026c), and even the numerical precision at which a small model was served changed how often it repeated the same brand (Kurtskhalia 2026). Prompt framing can also manufacture results: when prompts asked for the “most popular” or “best selling” colas, major brands took 56–68% of mentions (Chen et al 2025), an outcome the wording partly demands.

There are indications that instability is concentrated among less prominent brands. Persona effects were largest for mid-market brands in the full-text vendor audit (Jack et al 2026d), and in abstract-level studies category leaders kept about 80% same-brand consistency across personas while mid-market brands swapped up to three-quarters of their appearances (Jack et al 2026c), and distributed human auditors querying Google's AI Mode from six Indonesian islands agreed on a stable core of large national brands, which appeared for 97% of auditors, while disagreeing about less prominent names (Budiman 2026). This pattern rests on few independent groups and is reported here as tentative.

RQ3: Validity appraisal

Figure 2 summarises the appraisal. Summary scores ranged from 11% to 89% of the applicable maximum (median 44%). Three strengths stand out: most studies defined the brand universe independently of the model's outputs (22 fully, 10 partly), 16 took at least five draws per prompt and reported variance, and 15 released data, prompts or code. The weaknesses are systematic.

Criterion validity is rare. Only two studies related scores to an external brand or market benchmark: brand-equity rankings, advertising spend, search interest and online conversation (Malthouse et al 2026), and a firm-size benchmark (Žatuchin 2026h). Thirteen others used proxies such as stature tiers, prominence catalogues, domain-authority scores or census attributes such as review counts (Pitenin 2026). None related audit scores to outcomes managers care about, such as traffic, consideration or sales.

Time is rarely modelled. Twenty-nine studies measured once. The seven with repeated measurement show why it matters and how to do it: daily panels separated within-day noise from drift and recommended rolling multi-week windows (Schulte et al 2026; Chen and Koudas 2026), and a two-wave design found brand inclusion shares almost perfectly rank-stable across seven months (Žatuchin 2026h).

Prompt sampling is seldom justified. Five studies justified their prompts with a defensible sampling strategy, such as persona-based generation with manual verification, systematic paraphrase variation, or templates written natively or translated and back-translated across languages; one further study derived prompts from high-volume search keywords (Schulte et al 2026), the closest any came to sampling real buyer questions. Most used author-written or model-generated prompts without evidence that they represent buyer questions, which matters given the size of paraphrase effects.

Systems are underspecified. Ten studies reported model versions, access mode, retrieval setting, location and dates in full. Studies of consumer products often could not name the underlying model.

Brand extraction is rarely validated. In the 32 studies where brand names must be identified in free text, through alias dictionaries, substring matching or a second language model acting as judge, only two validated the extraction step against human coding (Žatuchin 2026c; Gonzalez Castro 2026), and six reported partial checks. Where it was measured, dictionary recall was 0.43 or lower (Žatuchin 2026c), and two LLM judges agreed at a Jaccard of only 0.65–0.67 (Jack et al 2026d). Extraction error is therefore a first-order source of measurement error, not a technical detail.

Sensitivity. Summary scores were higher for full-text than for abstract-level studies (mean 51% against 37%), partly because abstracts omit design details, and somewhat higher for studies without a commercial interest (49% against 43%). Repeating the conclusion counts with full-text studies only (Table 4, second column) leaves every family represented. The inclusion gradient (A1) and the stage-dependence of incumbent advantage rest mainly on full-text studies (Kumar 2026; Jack et al 2026d; Chu and Hou 2026; Sharma 2026). The documentation interpretation of prominence draws on full-text studies (Malthouse et al 2026; Chen and Koudas 2026; Sharma 2026) but its strongest evidence, the census-based separation of entry and position (Pitenin 2026), is abstract-level. The concentration of instability among less prominent brands depends substantially on abstract-level studies.

A two-stage account of brand visibility in machine answers

The synthesis suggests that many apparent contradictions in this literature arise because studies observe different parts of the answer-generation process, under different gates and through different routes. Figure 3 summarises a two-stage account; Table 5 states six propositions, the evidence behind each, an assessment of its current strength, and observations that would falsify it. The two stages are an analytic decomposition of observable outcomes, not a claim about how models compute answers.

The first stage is *membership*: whether a brand is included in the answer at all. This is the machine analogue of entry into the consideration set and, like recall-based entry (Nedungadi 1990), it appears to

depend on accessibility rather than evaluation. Open-generation audits converge on a steep gradient at this stage, and the near-binary pattern at the level of individual prompts (Kumar 2026) suggests that membership behaves less like a probability spread evenly across questions and more like a set of category entry points to which a brand is or is not linked, which is how salience is conceptualised for human memory (Romaniuk and Sharp 2004). The evidence further suggests that what drives membership is documented prominence, meaning the volume and reach of text about a brand that a model has learned from or can retrieve, rather than market position as such (P1).

The second stage is *selection*: which included brands are ranked first or recommended. Here the evidence is thinner and mixed. Attribute evidence can outweigh brand identity among supplied options (Chu and Hou 2026), ratings predicted first position but not entry in a restaurant census (Pitenin 2026), and conditional conversion to recommendation was not higher for leaders than for challengers in one audit (Jack et al 2026d). Yet brand names still shift rankings of otherwise similar products (Pfrommer et al 2024; Chen and Koudas 2026), and chain affiliation carried a choice premium for about half of the models in a hotel-booking benchmark (Kireyev 2026). P2 is therefore stated as a conjecture: the weight of prominence is smaller at selection than at membership, and attribute evidence matters more. If it holds, it mirrors the classic finding that consideration-set entry and choice within the set follow different logics (Hauser and Wernerfelt 1990; Shocker et al 1991), and it would qualify the expectation, carried over from recommender systems, that algorithmic intermediaries concentrate demand (Fleder and Hosanagar 2009): concentration, where it exists, would sit mainly at membership.

Two routes appear to lead into answers (P3). Prominent brands can enter through the model prior, that is, parametric associations that persist across sessions and that models write into their own search queries. Less prominent brands appear to depend more on retrieval, appearing when the system's searches find pages that name them. This suggests a view of brand salience for machine intermediaries with a stock component, the accumulated text that shapes the prior, and a flow component, the retrievable evidence available at query time, with the flow component mattering most for brands whose stock is small. It is also consistent with the finding that the effect of off-site signals differs between systems that retrieve differently (House 2026).

Before either stage operates, gates define the universe of eligible brands (P4). Query language, the market inferred from the user's location and the origin of the model shift which brands can be named, sometimes completely (Žatuchin 2026e, 2026f; Huang et al 2025). Visibility is therefore better treated as a property of a brand in a market, language and system than as a property of the brand.

Measurement noise may not be neutral across brands (P5): instability across draws, paraphrases, personas and providers appears to concentrate on less prominent brands, whose membership depends on fragile retrieval paths, while leaders form a stable core (Budiman 2026; Jack et al 2026c, d). Finally, the design of an audit determines which part of this process it observes (P6). Supplied lists observe selection only; open designs observe membership multiplied by selection; named-brand queries observe

representation. Prompt framing that invokes popularity selects for prominence by construction (Chen et al 2025).

Table 5
Propositions, supporting evidence, current strength and falsifying observations

Proposition	Principal evidence	Strength of current evidence	Observation that would falsify it
P1 Membership. In open-generation answers, inclusion rises steeply with a brand's documented prominence (volume and reach of text about it) and is near-binary at the prompt level; it tracks documentation more closely than market position.	Kumar (2026); Jack et al (2026d); Sharma (2026); Malthouse et al (2026); Żatuchin (2026h); House (2026); Pitenin (2026); Chen et al (2026)	Moderate for the gradient (several groups, full text); emerging for documentation versus market position	Inclusion rates flat across documentation tiers, or more closely aligned with market share than with documentation, under externally defined brand sets and repeated draws
P2 Selection (conjecture). Conditional on inclusion, the weight of prominence in being recommended or ranked first is smaller than at membership, and attribute evidence matters more.	Chu and Hou (2026); Pitenin (2026); Jack et al (2026d); contrary: Pfrommer et al (2024); Chen and Koudas (2026); Kireyev (2026)	Weak and mixed; conditional estimates are affected by selection on inclusion	Prominence effects on conditional selection as large as on inclusion in designs that hold query–brand fit constant
P3 Two routes. Prominent brands can enter through the model prior; less prominent brands depend more on retrieved third-party and own-domain evidence.	Chen and Koudas (2026); Tannenbaum (2026); Jack et al (2026a); Hamann et al (2026); House (2026); Sharma (2026)	Moderate, observational only	Equal sensitivity of leaders and challengers to removal of retrievable evidence in a controlled design
P4 Gates. Query language, inferred market and model origin define the eligible brand universe before membership.	Żatuchin (2026e, 2026f, 2026g); Huang et al (2025); Rienecker et al (2026); Kamruzzaman et al (2024)	Moderate; concentrated in one group for the language effect	Brand sets invariant to query language and location for brands with strong home markets
P5 Tail instability. Variance across draws, paraphrases, personas, providers and time concentrates on less prominent brands.	Jack et al (2026c, d); Budiman (2026)	Tentative; partly abstract-level	Reliability of visibility scores that does not decline with prominence
P6 Design lens. Measured incumbent advantage depends on paradigm: supplied lists show it mainly at attribute parity; open designs show it strongly; popularity-framed prompts inflate it.	Chu and Hou (2026) versus Kumar (2026) and Jack et al (2026d); Chen et al (2025)	Moderate, by cross-study comparison only	Equal incumbent advantage across supplied and open designs applied to the same brands and categories

A minimum design and reporting standard

The appraisal and the framework jointly imply a minimum standard for any audit that reports brand visibility in LLM answers, whether academic or commercial (Table 6). Several items echo recommendations made within individual studies, for example on repeated runs (Schulte et al 2026; Žatuchin 2026d), consideration-set metrics (Malthouse et al 2026), prompt corpora (Martinez 2026b) and separable measures (Khan 2026); the contribution here is to integrate them with criteria that the reviewed evidence shows are rarely met, notably criterion validity, temporal design and validated brand extraction. Each item is achievable with public interfaces and modest budgets. Reporting membership and conditional selection separately (item 1) could help resolve many of the contradictions identified in this review. Validating brand extraction (item 5) is cheap relative to its consequences, given that an unvalidated matcher can change which brand appears to lead a category.

Table 6
Minimum design and reporting standard for brand-visibility audits

#	Standard	Minimum practice	Criteria
1	Name the construct and stage	Report membership (inclusion), conditional selection (recommended or top-ranked given inclusion), representation (what is said) and sourcing as separate outcomes with explicit denominators	V1
2	Fix the market frame	State language(s), user location or IP, and market; audit each relevant language for brands with home markets outside English	V3, V5
3	Sample prompts from buyer questions	Derive intents from documented buyer language (search queries, category entry points, reviews); use several paraphrases per intent; report results per intent and paraphrase variance	V3
4	Take enough draws and report uncertainty	Pilot to estimate variance by facet, then set draws per prompt (the reviewed studies suggest at least 5–10 per prompt per occasion); report intervals clustered by prompt	V2
5	Define brands externally and validate extraction	List brands and aliases in advance from external sources while also recording unlisted brands; report precision and recall of extraction against a human-coded sample	V4, V9
6	Specify the system	Provider, model version, consumer product or programming interface, retrieval setting, date and time, temperature where settable; log retrieved sources	V5
7	Validate against external criteria	Relate scores to at least one external brand or market metric (market share, search interest, tracked awareness) and, where possible, to traffic or sales	V6
8	Measure over time	Repeat on a schedule; separate within-occasion noise from drift with rolling windows; record model updates	V7
9	Be transparent about data and interests	Release prompts, raw responses and code; declare commercial interests, including interests in measured brands	V8

Discussion

Theoretical implications

The review contributes to marketing theory in three ways. First, it extends consideration-set theory to machine intermediaries. Separating the inclusion of a brand in an answer from its ranking within the answer reconciles a substantial share of findings that otherwise look contradictory, in the same way that separating entry from choice clarified human decision-making (Hauser and Wernerfelt 1990; Shocker et al 1991). The answer produced by an LLM can be studied as a *machine consideration set* with its own entry rules, partly inherited from the text the model has learned from and partly determined at query time by retrieval.

Second, it refines the idea of brand salience for an environment in which machines, not only people, must “think of” a brand. Mental availability depends on links to many category entry points (Romaniuk and Sharp 2004); its machine counterpart appears to depend on how extensively a brand is documented in the text available to the model, both in training data and at retrieval time. This complements proposals that LLMs construct rather than merely reflect brand knowledge (Hamann et al 2026) and measures that separate parametric recall from inference-time activation (Lüders 2026). It also implies that marketing investments which build human salience without leaving a durable textual footprint may not transfer to machine answers, whereas documentation that consumers rarely read, such as directory listings and reviews, may matter disproportionately.

Third, the review shows that measurement theory, not only optimisation, belongs at the centre of research on generative AI and brands. Agenda-setting work calls for marketing to study AI agents as customers (Sarstedt et al 2026). The present evidence indicates that before such work can estimate effects, it needs construct definitions, sampling frames and reliability standards adapted to stochastic, market-specific and rapidly changing systems. Table 6 offers a starting point.

Managerial implications

For managers, the most immediate implication concerns how to read AI-visibility dashboards. A single visibility percentage should prompt five questions: which stage does it measure, over which prompts and languages, with how many draws per prompt, with what uncertainty, and how were brand names identified? A score without answers to these questions is a weak basis for budget decisions.

Where a brand is constrained should guide strategy, with the caveat that the supporting evidence is observational. For less prominent brands the binding constraint appears to be membership, and the evidence points to documentation that answer engines retrieve: being named on third-party pages, in reference sites, directories, review platforms and communities, and having crawlable own-domain content with basic facts such as prices (Chen and Koudas 2026; House 2026; Tannenbaum 2026; Pitenin 2026). On-page features marketed as GEO explained little variation in citations within websites in the one study that compared them with retrieval (Chen and Koudas 2026), and a composite GEO score was unrelated to discovery (Sharma 2026). Aggressive manipulation is risky: an injected instruction promoted a brand in one model family but demoted it in another (Paeng 2026). For leading brands, membership is largely secure, and attention can shift to selection and representation: whether the attribute evidence available to models supports recommendation, and whether what is said is accurate and current (Gonzalez Castro 2026; Varga 2026). Brands with strong home markets should audit in their customers’ languages, because English-language measurement can understate them (Žatuchin 2026e).

Policy implications

Concerns that generative AI will entrench dominant brands need a precise target. The evidence reviewed here places the clearest concentration in which brands are named at all; whether named brands are favoured in recommendation is less clear. Competition and consumer-protection analysis may therefore benefit from examining membership: which brands are eligible for answers, how retrieval sources are

chosen, and whether commercial arrangements influence either. The prevalence of studies by authors with commercial stakes, and the absence of standard reporting, also suggest a role for disclosure norms in the marketing of AI-visibility metrics.

Limitations

The review has limitations that follow from the field and from its design. The search relied on open sources and could not include subscription business databases; Google Scholar was screened only to the first results per string, forward citation searching was incomplete, and some relevant work, including a peer-reviewed study of brand-origin bias in e-commerce rankings and a longitudinal industry analysis, could not be retrieved. Screening and charting were performed by an LLM, and reliability was checked with a blind second LLM screener on a 20% random sample ($\kappa = 0.81$) rather than with a human second reviewer; this design may miss or misclassify studies, and the coding of conclusion families was inductive. Thirteen studies were charted from abstracts and thirteen through structured extraction of targeted sections; both are flagged, quoted figures were checked, and a sensitivity check is reported, but errors may remain. The protocol was written on the day of the searches and was not registered. The evidence base itself is dominated by unrefereed and commercially interested work and is clustered in a few groups, so the number of papers overstates the number of independent observations. The systems studied change faster than the literature, so specific estimates will date quickly even if the structural patterns do not. Finally, the propositions are derived from a synthesis of mostly observational audits; they are claims to be tested, not established effects.

Future research

Six directions follow. (1) *Criterion validation*: relate audit scores to brand-tracking measures of awareness and consideration, to search and traffic data, and ultimately to sales, for the same brands and markets. (2) *Stage-separated designs*: report membership and conditional selection for the same brands under open and supplied paradigms, holding query–brand fit constant, to test P1, P2 and P6 directly. (3) *Route experiments*: vary the retrievable evidence about matched brands of differing prominence in controlled settings to test P3. (4) *Consumer-product panels across markets*: track the products buyers use rather than programming interfaces, across languages and locations, to estimate gates (P4) and drift. (5) *Category moderators*: test whether category maturity, involvement, or hedonic versus utilitarian benefits moderate membership and selection, as the higher agreement observed in consolidated categories suggests (Budiman 2026). (6) *Independent replication*: re-run influential vendor studies with shared, preregistered protocols and open data. As AI agents begin to complete purchases rather than only advise, these questions will move from measurement to market outcomes.

Conclusion

Brand-visibility audits of LLMs have grown faster than the standards needed to interpret them. Across 38 studies, measurement is fragile: few audits validate against external brand or market benchmarks, track change over time, or check how they identify brands in text, and much of the evidence comes from a few commercially interested groups. The evidence nonetheless suggests a coherent picture that future work can test. How extensively a brand is documented online appears to shape whether it is named; attribute evidence may matter more for which named brand is recommended; and language, market and system decide which brands are eligible in the first place. Separating these stages, and measuring each to the standard marketing applies to any brand metric, will let researchers test how generative AI is reshaping brand competition and help managers judge which of their dashboards to trust.

Declarations

Use of generative AI. The use of a large language model in conducting this review and drafting the manuscript is described in the Method section (Use of generative AI in conducting the review). The model is not an author, and the author takes full responsibility for the content.

Competing interests.

Declared on the separate title page to preserve anonymity during review.

Funding.

The author received no funding for this work.

Data availability.

The search log, screening log, charting data, validity scores and decision log are provided as an open evidence register (Online Appendices A and B) and will be deposited in a public repository with a persistent identifier on acceptance.

Ethics.

The study analyses published and publicly posted research and did not involve human participants or personal data.

References

1. Aggarwal P, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K, Deshpande A (2024) GEO: Generative engine optimization. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery

- and Data Mining, 5–16. <https://doi.org/10.1145/3637528.3671900>
2. Arksey H, O'Malley L (2005) Scoping studies: Towards a methodological framework. *Int J Soc Res Methodol* 8(1):19–32. <https://doi.org/10.1080/1364557032000119616>
 3. Arshad M (2026) Cited or invisible? A multilingual, multi-platform audit of brand citation in generative search. Technical report. Zenodo. <https://doi.org/10.5281/zenodo.22491737>
 4. Budiman D (2026) Auditing AI together: A distributed human-auditor study of brand convergence and citation structure in Google AI Mode. SSRN Working Paper 6972038. <https://doi.org/10.2139/ssrn.6972038>
 5. Chen L, Weng G, Moro E (2026) Large language models create an uneven informational layer over cities. arXiv preprint. <https://doi.org/10.48550/arXiv.2607.06260>. arXiv:2607.06260
 6. Chen M, Koudas N (2026) What drives AI citations? Content, retrieval and model knowledge in generative engine optimization (GEO). arXiv preprint arXiv:2610.15033. <https://doi.org/10.48550/arXiv.2610.15033>
 7. Chen M, Wang X, Chen K, Koudas N (2025) Generative engine optimization: How to dominate AI search. arXiv preprint. <https://doi.org/10.48550/arXiv.2509.08919>. arXiv:2509.08919
 8. Chu X, Hou Y (2026) Incumbent advantage: Brand bias and cognitive manipulation dynamics in LLM recommendation systems. arXiv preprint arXiv. <https://doi.org/10.48550/arXiv.2606.17443>. :2606.17443
 9. Churchill GA Jr (1979) A paradigm for developing better measures of marketing constructs. *J Mark Res* 16(1):64–73. <https://doi.org/10.1177/002224377901600110>
 10. Cook N (2025) Mind the invisibility gap: An analysis of 80 leading global brands in LLM answers. SSRN Working Paper 5698748. <https://doi.org/10.2139/ssrn.5698748>
 11. Ehrenberg ASC, Goodhardt GJ, Barwise TP (1990) Double jeopardy revisited. *J Mark* 54(3):82–91. <https://doi.org/10.1177/002224299005400307>
 12. Filandrianos G, Dimitriou A, Lymperaio M, Thomas K, Stamou G (2025) Bias beware: The impact of cognitive biases on LLM-driven product recommendations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 22397–22426. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1140>
 13. Fleder D, Hosanagar K (2009) Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Manage Sci* 55(5):697–712. <https://doi.org/10.1287/mnsc.1080.0974>
 14. Gonzalez Castro PU (2026) A designed-for-test measurement of phantom-brand presence in large language model outputs. SSRN Working Paper 6721779. <https://doi.org/10.2139/ssrn.6721779>
 15. Hamann A, Stahl F, Kannan PK (2026) Beyond consumer memory: How large language models update brand knowledge. SSRN Working Paper 6798479. <https://doi.org/10.2139/ssrn.6798479>
 16. Hannak A, Sapiezynski P, Molavi Kakhki A, Krishnamurthy B, Lazer D, Mislove A, Wilson C (2013) Measuring personalization of web search. In *Proceedings of the 22nd International Conference on World Wide Web*, 527–538. <https://doi.org/10.1145/2488388.2488435>

17. Hauser JR, Wernerfelt B (1990) An evaluation cost model of consideration sets. *J Consum Res* 16(4):393–408. <https://doi.org/10.1086/209225>
18. House J (2026) Off-page signals have model-specific effects on generative AI search visibility: Evidence from a cross-platform audit of 2,729 businesses across five generative AI systems. SSRN Working Paper 6731878. <https://doi.org/10.2139/ssrn.6731878>
19. Huang J, Situ R, Ye R (2025) Cultural encoding in large language models: The existence gap in AI-mediated brand discovery. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2601.00869>. arXiv:2601.00869
20. Iannelli M, Ai A (2026) From prompt to purchase: How AI brand recommendations move consumers on the open web. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2606.10907>. arXiv:2606.10907
21. Jack W, Lehman N, Maloney K, Xu S (2026a) Divergent recommendations, convergent diagnoses: Cross-provider failure-mode convergence in AI commercial recommendation. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2606.26116>. arXiv:2606.26116
22. Jack W, Lehman N, Maloney K, Xu S (2026b) Paraphrase brittleness in production retrieval-augmented commercial recommendation: Reproducibility below the rerun-stability baseline. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2605.27440>. arXiv:2605.27440
23. Jack W, Lehman N, Maloney K, Xu S (2026c) Persona conditioning of brand recommendations in retrieval-augmented commercial chat: A prominence-stratified cross-provider audit. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2605.30207>. arXiv:2605.30207
24. Jack W, Lehman N, Maloney K, Xu S (2026d) Prominence-stratified failure modes in retrieval-augmented commercial recommendation: A 37,000-run audit. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2605.27439>. arXiv:2605.27439
25. Kamruzzaman M, Nguyen HM, Kim GL (2024) Global is good, local is bad? Understanding brand bias in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12695–12702. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.707>
26. Keller KL (1993) Conceptualizing, measuring, and managing customer-based brand equity. *J Mark* 57(1):1–22. <https://doi.org/10.1177/002224299305700101>
27. Khan B (2026) Brand representation in AI answers: An open specification for measuring visibility, alignment and sentiment. SSRN Working Paper 6971558. <https://doi.org/10.2139/ssrn.6971558>
28. Kireyev P (2026) PriceBench: A diagnostic benchmark for price, quality, and brand preferences in LLM booking agents. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2609.31468>. arXiv:2609.31468
29. Kumar P (2026) Generative engine optimization at scale: Measuring brand visibility across AI search engines. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2606.20065>. arXiv:2606.20065
30. Kurtskhalia D (2026) Quantization amplifies determinism, not bias: Scale-dependent behavioral effects of serving-time weight compression. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2609.07901>. arXiv:2609.07901

31. Lüders S (2026) Measuring mental availability in the generative era: Introducing the Generative Visibility Score (GVS) – A deterministic framework for brand salience in large language models. SSRN Working Paper 6567201. <https://doi.org/10.2139/ssrn.6567201>
32. MacKenzie SB, Podsakoff PM, Podsakoff NP (2011) Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Q* 35(2):293–334. <https://doi.org/10.2307/23044045>
33. Malthouse E, Lee K-Y, Yang J, Pal S, Feng X (2026) Evaluating brand retrieval and ranking in large language model recommendations. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2609.16304>. arXiv:2609.16304
34. Martinez O (2026a) Optimizing visibility in generative engines: A critical survey of generative engine optimization (2023–2026). <https://doi.org/10.48550/arXiv.2607.14035>. *arXiv preprint* arXiv:2607.14035
35. Martinez O (2026b) Measuring GEO visibility: Prompt corpora define the answer market. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2609.06811>. arXiv:2609.06811
36. Messick S (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 50(9):741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
37. Metaxa D, Park JS, Robertson RE, Karahalios K, Wilson C, Hancock J, Sandvig C (2021) Auditing algorithms: Understanding algorithmic systems from the outside in. *Found Trends Human–Computer Interact* 14(4):272–344. <https://doi.org/10.1561/11000000083>
38. Nedungadi P (1990) Recall and consumer consideration sets: Influencing choice without altering brand evaluations. *J Consum Res* 17(3):263–276. <https://doi.org/10.1086/208556>
39. Paeng H (2026) The injection paradox: Brand-level suppression in safety-trained LLM recommendations via RAG context injection. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2606.09204>. arXiv:2606.09204
40. Pfrommer S, Bai Y, Gautam T, Sojoudi S (2024) Ranking manipulation for conversational search engines. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9523–9552. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.534>
41. Pitenin V (2026) Invisible to the machine: Auditing AI restaurant, cafe, and bar recommendation against a complete market census. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2608.07069>. arXiv:2608.07069
42. Rienecker J, Mpofu K, Goel N, Datta S, Zhao J, Danielsson O, Thorsen F (2026) Auditing preferences for brands and cultures in LLMs. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2603.18300>. arXiv:2603.18300
43. Romaniuk J, Sharp B (2004) Conceptualizing and measuring brand salience. *Mark Theory* 4(4):327–342. <https://doi.org/10.1177/1470593104047643>

44. Sarstedt M, Adler SJ, Imschloss M (2026) Machine marketing: Rethinking the customer in the age of generative AI. *J Mark Analytics*. Advance online publication <https://doi.org/10.1057/s41270-026-00521-y>
45. Schulte J, Bleeker M, Kaufmann P (2026) Don't measure once: Measuring visibility in AI search (GEO). *arXiv preprint*. <https://doi.org/10.48550/arXiv.2604.07585>. arXiv:2604.07585
46. Sharma AP (2026) The discovery gap: How Product Hunt startups vanish in LLM organic discovery queries. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2601.00912>. arXiv:2601.00912
47. Shocker AD, Ben-Akiva M, Boccara B, Nedungadi P (1991) Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Mark Lett* 2(3):181–197. <https://doi.org/10.1007/BF02404071>
48. Tannenbaum B (2026) From prompt to recommendation: A fitted stage model of brand visibility in AI search. *arXiv preprint* arXiv:2609.23162. <https://doi.org/10.48550/arXiv.2609.23162>
49. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D et al (2018) PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann Intern Med* 169(7):467–473. <https://doi.org/10.7326/M18-0850>
50. Varga Z (2026) Per-entity bias mapping for AI visibility: Why brand mentions require entity-specific calibration. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2606.21595>. arXiv:2606.21595
51. Zac A, Gal MS (2025) The price of advice: Experimental evidence on the effects of AI recommenders. *New Working Paper Series No. 375*. Chicago: Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business. <https://hdl.handle.net/10419/336742>
52. Žatuchin D (2026a) Gender bias in large language model brand recommendations: A three-study analysis of prompt-induced disparities across seasonal and recipient contexts. *Res Square preprint*. <https://doi.org/10.21203/rs.3.rs-8883056/v1>
53. Žatuchin D (2026b) How large language models source brand reputation across languages and markets. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2606.25787>. arXiv:2606.25787
54. Žatuchin D (2026c) System attribution in LLM brand recommendations: Single responses identify the system, aggregated brand profiles do not transfer. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2610.00253>. arXiv:2610.00253
55. Žatuchin D (2026d) The Dice Roll Method: A standardized protocol for repeated-query auditing of large language model brand recommendations. *arXiv preprint* arXiv:2609.04047 (v2). <https://doi.org/10.48550/arXiv.2609.04047>
56. Žatuchin D (2026e) The language blind spot: How query language and brand recognition tier shape AI-constructed brand reputation across twelve European languages. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2606.23165>. arXiv:2606.23165
57. Žatuchin D (2026f) The language of the question selects the market: Query language and exit IP as separable factors in commercial recommendations from a generative search interface. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2608.30052>. arXiv:2608.30052

58. Žatuchin D (2026g) Where does the noise come from? A variance-components decomposition of non-determinism in LLM brand answers. arXiv preprint. <https://doi.org/10.48550/arXiv.2607.13304>. arXiv:2607.13304
59. Žatuchin D (2026h) Who owns the AI recommendation? A multi-industry empirical map of brand category ownership across large language models. arXiv preprint arXiv:2606.23057 (v2). <https://doi.org/10.48550/arXiv.2606.23057>

Figures

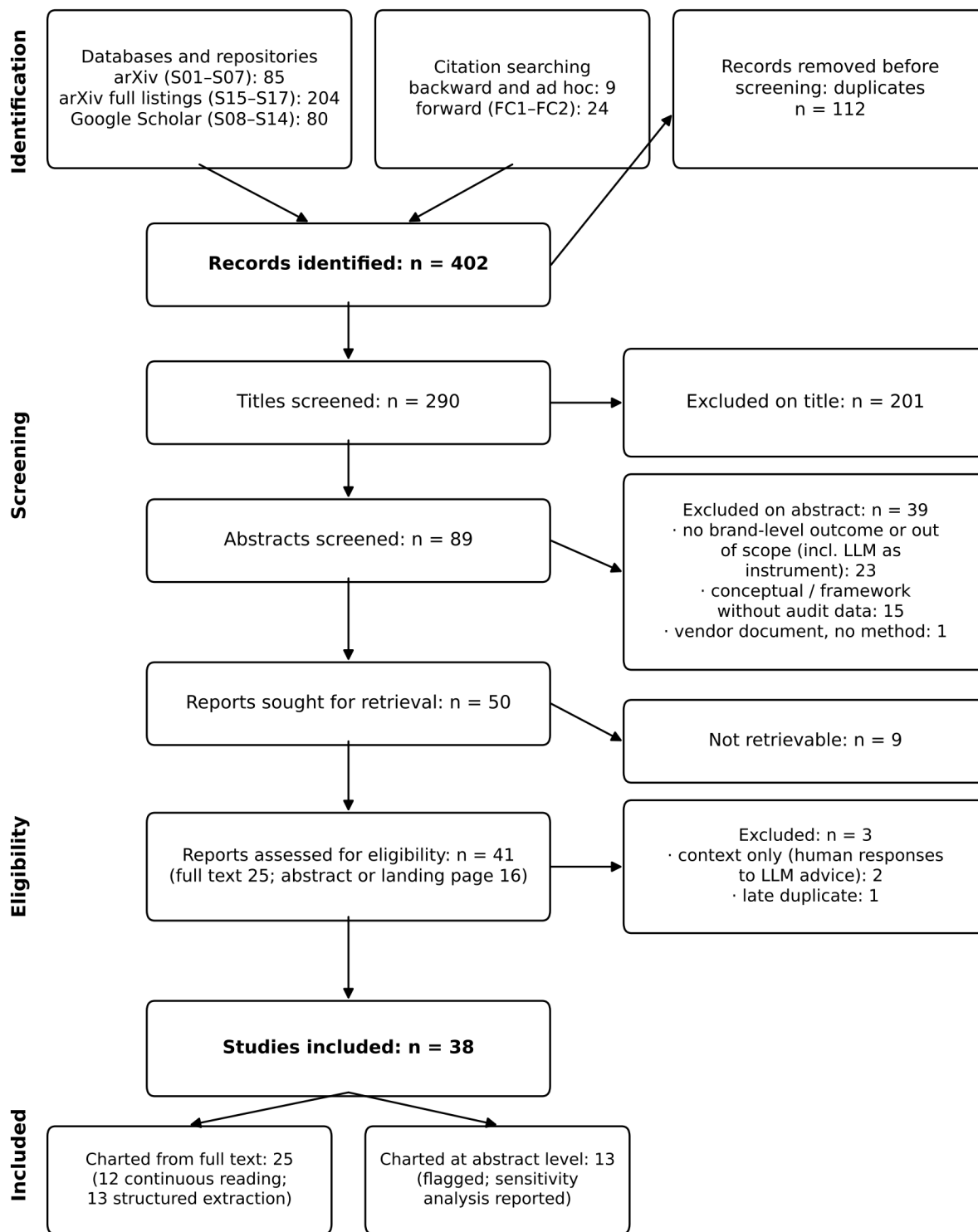


Figure 1

Flow of records through the review (PRISMA-ScR)

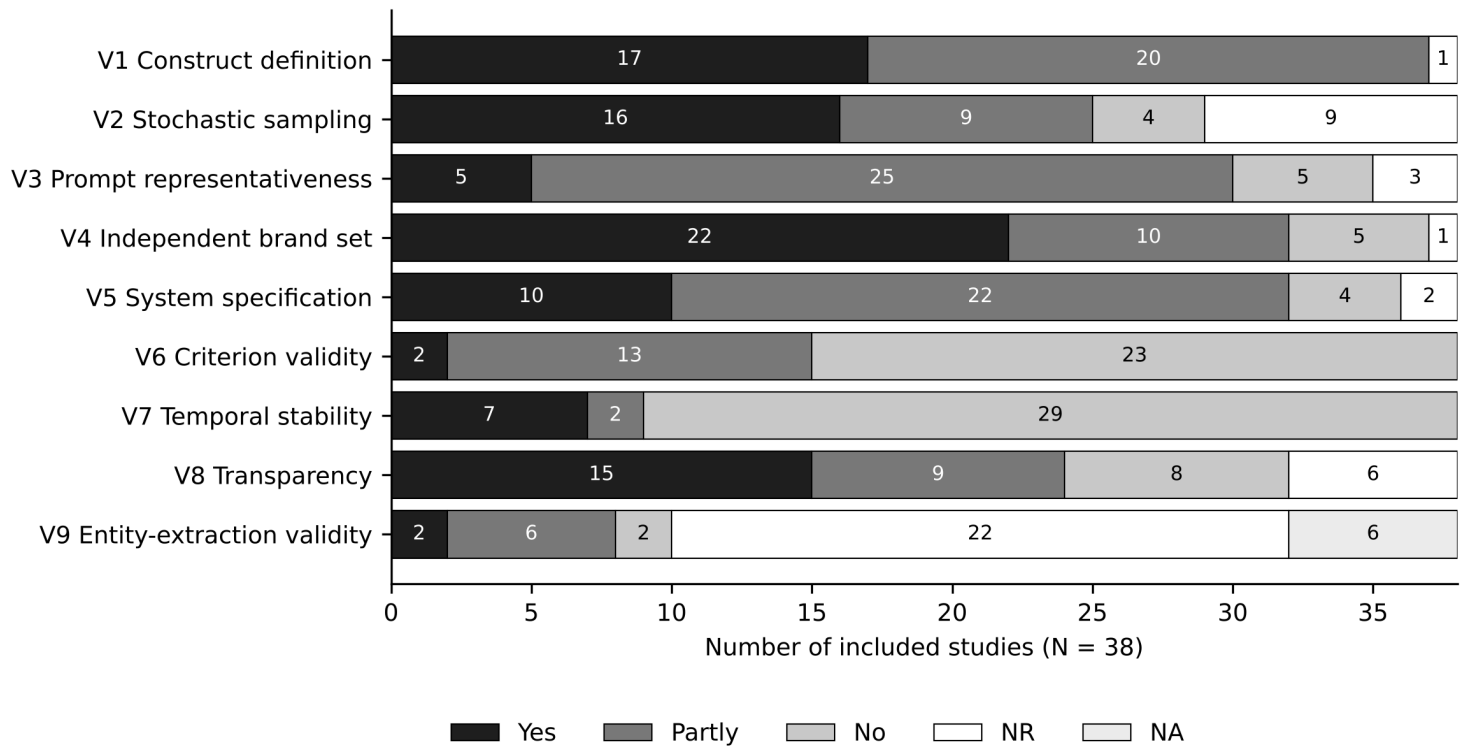


Figure 2

Validity appraisal of the 38 included studies on nine criteria (NR = not reported; NA = not applicable)

P5: instability across runs, paraphrases, personas and providers appears concentrated among less prominent brands

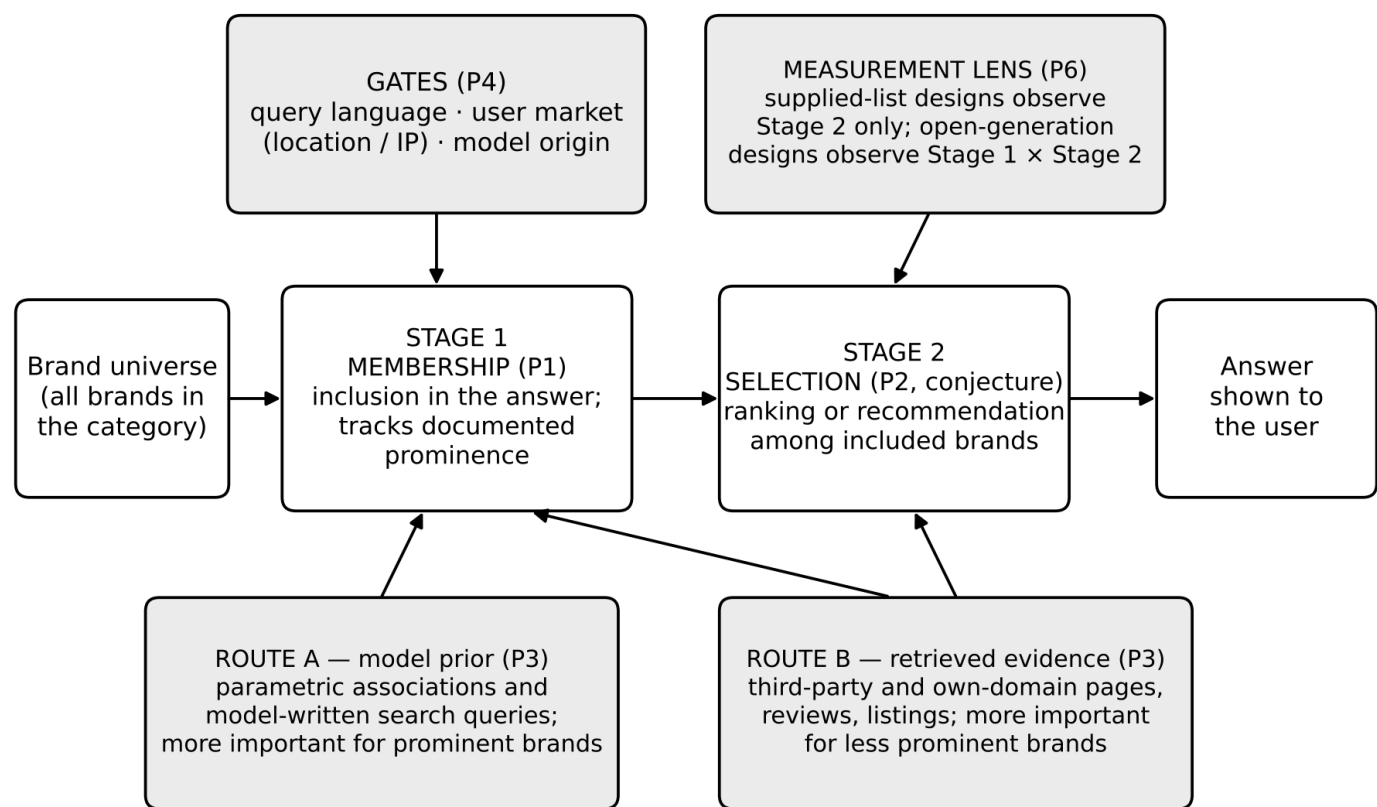


Figure 3

A two-stage account of brand visibility in LLM answers. P1–P6 refer to the propositions in Table 5

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryAMethodologyAppendix.docx](#)
- [SupplementaryBEvidenceRegister.xlsx](#)