AI VISIBILITY · MEASUREMENT · GBMF · TACTICAL GUIDE

How to measure your brand's AI visibility: a GBMF walkthrough

A practitioner walkthrough for running your first GBMF measurement. Prompt set, engine set, Brand Profile, evaluator, and how to read the output.

AIVIARA RESEARCH · MAY 2026

You have read the framework explainer and want to run your own measurement. This article walks through that as a practitioner: what decisions to make before you start, what an operational first scan looks like, and how to read the output. The companion explainer on the three measures covers what GBMF is and why; this is the operational follow-on.

The simplified pipeline below is enough to run a first internal scan and surface where the brand sits on the visibility-alignment grid. The full methodology (formal definitions, error quantification, evaluator architecture, falsifiability thresholds) lives in the GBMF Working Paper; the long-form practitioner guide covers the production-grade version. This article is the version a marketing or brand-monitoring lead can run before either of those.

The four decisions before any scan runs

A GBMF measurement is defined by four things. None is optional; all four go into the conditions block that gets reported alongside the score, so other parties can interpret the result correctly.

The prompt set. A category-level question set the practitioner believes a target buyer would actually ask. Three query types matter: category discovery ("best analytics software for B2B mid-market"), brand-direct ("does Acme Analytics integrate with Salesforce"), and use-case ("how do I track usage-based pricing across departments"). The GBMF paper recommends 30 prompts per category for benchmark publication; a first scan can start with 10 to 15 and expand from there.

The engine set. A declared list of AI engines the scan runs against. The standard list at time of writing covers ChatGPT, Claude, Microsoft Copilot, Gemini, Google AI Overviews, and

Perplexity, encoded as AIO-CLD-COP-GEM-GPT-PPX. The point of declaring the list is reproducibility; two studies are not comparable unless they ran against the same engines.

The Brand Profile. A version-controlled document of the brand's own declared facts. Pricing entry point, primary product category, founding year, HQ country, plus claimed features and positioning. MA is scored against this profile.

The evaluator approach. How MA and MS will be scored from the collected responses. Three options exist in practice. An LLM-as-judge approach (cross-family is required for credibility) scales to thousands of responses but carries known biases. A human gold standard produces higher-confidence scoring on a smaller sample at higher cost. A hybrid runs the LLM at scale with human calibration against a 200-response sample.

Designing the prompt set

The full methodology covers prompt-set construction in Appendix A of the GBMF paper. The simplified version for a first scan:

Cover the three query types in roughly equal proportion. Skew the mix toward where buyers actually ask questions, not where the brand wishes they did. Use buyer interviews, support transcripts, or sales-call recordings to source the prompts; the framework's whole logic depends on the prompts being representative of buyer questions, not internal positioning.

Avoid brand names from the competitive set in category-discovery and use-case prompts. Brand-direct prompts name only the brand they target. Version-lock the prompt set as soon as the scan starts. Mid-scan edits invalidate the comparison. And run each prompt at least three times per engine. Single-run measurements are unreliable under probabilistic response variance (Schulte, Bleeker & Kaufmann, 2026, arXiv:2604.07585; preprint).

A common first-time error is to build the prompt set around terms the brand uses internally rather than terms buyers actually search with. The brand "wins" the category benchmark by construction, and the result is uninformative.

Weak prompt (internal-positioning):

> **"Which AI-native usage analytics platform offers the most flexible attribution model for SaaS metering?"**

Strong prompt (buyer-question):

> "Best tools for tracking usage-based billing across product features for B2B SaaS"

The first prompt embeds terminology only a vendor would use. The second describes a buyer goal. AI engines respond to the second with the candidate set buyers actually see.

Choosing the engine set

The framework requires a declared engine set rather than a single engine because cross-engine fragmentation is large. The Digital Bloom's December 2025 synthesis of 680 million citations found that only 11% of sites are cited by both ChatGPT and Perplexity. The two platforms draw from substantially different source pools, and a single-engine scan systematically underrepresents a brand's actual AI footprint.

Chatoptic's September 2025 controlled study (Omer Ben-Porat at the Technion, 15 brands × 1,000 queries) found that Google rank is a poor predictor of ChatGPT citation: Spearman correlation of 0.034 with browsing enabled, 0.022 without. Two systems with different objectives. Running only against the engine you happen to use the most produces a number that is not interpretable as your brand's AI footprint.

A first scan should cover the standard six. Score each engine independently, then compute the headline as the equal-weighted mean over engines clearing the eligibility floor. Per-engine reporting is required, not optional. Aggregating away the per-engine numbers hides the cross-engine fragmentation the headline is averaging over.

Building a Brand Profile

The Brand Profile is the reference standard MA is scored against. For a B2B SaaS company doing a first scan, a minimum-viable profile looks like this:

Field	Type	Example for Acme Analytics
Primary product category	Tier 1 (externally verifiable)	"B2B analytics and reporting software"
Pricing entry point	Tier 1	"$89/month"
Founding year	Tier 1	"2019"

HQ country	Tier 1	"United Kingdom"
Key claim 1	Tier 2 (brand-stated)	"Integrates with all major CRM platforms"
Key claim 2	Tier 2	"No-code dashboard builder"
Disambiguation	Tier 2 (when relevant)	"Acme Analytics is not a CRM; it does not store contact data"

The Tier 1 minimum is three externally verifiable fields. The disambiguation field is the one teams most often skip and most often regret. When a brand name has any chance of being confused with another (a common brand name, a previous use of the name, a similarly-named competitor), the absence of a disambiguation statement can mean MA scoring credits the wrong entity entirely.

Watch out for profiles that read like marketing brochures. The MA score reflects the gap between what AI says and what the profile claims. An aspirational profile ("we are the leading provider of...") produces a deflated MA score, because AI describes the brand as it currently exists, not as the brand would like to be described. The profile should describe the brand as it is.

Scoring MA and MS

A practitioner has three operational options for scoring once the responses are collected.

The LLM-as-judge approach runs an evaluator LLM against the collected responses, scoring each one against the rubric in the GBMF paper. Cross-family is essential: if the production answer comes from a GPT-family engine, the evaluator must not be from the same family, because same-family evaluators inflate that engine's score relative to others (the family preference is part of the bias literature on LLM-as-judge, including Zheng et al. 2023 on position, verbosity, self-preference, and prompt-sensitivity biases). This approach scales but requires calibration against a small human sample.

The human gold standard approach has two or three trained annotators score a 200-response stratified sample directly. Trained-annotator scoring on a 200-response sample for a single category sits in the £3,500–£4,500 range at current Prolific rates. Higher confidence, slower, and the right approach when the scan is going to anchor public reporting.

The hybrid approach is what the production methodology recommends. The LLM evaluator scores at scale; a 200-response human gold standard licenses the evaluator for production

use against published agreement targets (Krippendorff's α ≥ 0.75 between the LLM and adjudicated humans). For a first scan on 10 to 15 prompts × 5 engines × 3 runs (around 225 responses), human-only scoring is feasible. As soon as the prompt set grows past about 30 prompts, the hybrid approach becomes necessary.

Another failure mode worth naming concerns evaluator selection. Using a single evaluator from the same family as one of the engines being scored inflates that engine's score and depresses the others, producing per-engine numbers that read as engine-level differences but reflect the evaluator's family bias.

Reading the output

The framework's headline output for a brand is three numbers (MV, MA, MS net) plus the sentiment distribution (positive / neutral / negative shares) and the state label that follows.

A worked example from the GBMF Working Paper: Acme Analytics scores MV 52, MA 67, MS +25 across a five-engine scan. The sentiment distribution is 51% positive, 23% neutral, 26% negative. MV ≥ 50 (visible), MA ≥ 50 (aligned), MS bands as Mixed because both positive and negative shares clear 25%. The state is *AI Wildcard*. The brand is on the buyer's shortlist as the divisive option, praised for capability and criticised on price.

Per-engine reporting goes alongside the headline. If GPT and Claude produce MV 62 and 58 while Copilot produces MV 40, the headline of 52 hides a substantive cross-engine gap. The headline is the equal-weighted mean of the five per-engine MVs; the engine-specific numbers carry the diagnostic information.

The state label is a presentation convention derived from the scores, not a separately validated construct. What gives it operational meaning is that each state implies a different remediation. The remediation by state article walks through what to do for each off-target state.

What this playbook will not achieve

A first scan tells you where you sit. It does not predict purchase. The framework measures the answer space, not the buyer's downstream decision. A brand can score *AI Champion* across the engine set and still lose deals for reasons GBMF does not measure. The state is a diagnostic input, not a sales forecast.

A first scan is not benchmark-grade. The GBMF paper sets a publishable-benchmark floor at $n \times k \geq 150$ responses per engine, which means 30 prompts × 5 runs or 50 × 3. A first scan beneath that floor is a useful internal diagnostic and should be reported as indicative; the conditions block carries the scan parameters so the indicative label is explicit.

A scan is conditional on the prompt set and engine set. Reports that omit the conditions block are not interpretable or comparable to other reports. The same brand measured against two different prompt sets is reporting on two different things.

A scan does not produce remediation. It tells you which state the brand is in; what to do about that state is a separate exercise, covered in the next article in this series.

The framework does not capture every form of AI visibility. Brands cited only as source URL footnotes, without being named in the response text, do not register as mentions in this framework. That is a deliberate boundary. The question GBMF asks is whether the brand exists in the answer space the buyer reads, not whether its content was used as a source.

Aiviara is building infrastructure for monitoring AI brand citations and factual accuracy across LLM platforms. Early access information is available at aiviara.com.