The Generative Brand Mention Framework: A Practitioner's Guide

Measuring brand visibility, alignment, and sentiment in AI-generated responses

Aiviara Research

June 2026

Executive Summary

AI systems are becoming a growing channel through which buyers discover, evaluate, and shortlist brands. Yet no standardised measurement infrastructure exists for what those systems actually say. Brands have invested heavily in managing their search visibility and have decades of tooling to support it. For AI, that infrastructure does not yet exist.

This guide introduces the Generative Brand Mention Framework (GBMF), an open measurement specification for brand representation in AI-generated responses. GBMF comprises three measures. Each addresses a separable property of how a brand is represented when an AI system answers a category-relevant question.

The first, MV (brand-mention visibility), measures how consistently AI systems mention a brand across a standardised set of prompts. It is reported on a 0–100 scale. A brand with an MV of 52 appears in an estimated 52% of AI responses to the prompts in its declared set, averaged across the declared engines.

The second, MA (brand-mention alignment), measures how accurately those mentions describe the brand. It is scored against a structured document called the Brand Profile, which the brand provides and which is version-controlled. MA is reported on a 0–100 scale. An MA of 67 means the average eligible mentioning cell receives an alignment score of 67/100 against the Brand Profile.

The third, MS (brand-mention sentiment), measures whether AI responses speak positively or negatively about the brand. MS is reported as three percentage shares (positive, neutral, negative) plus a net figure on a signed −100 to +100 scale. A net of +25 means positive characterisations outnumber negative ones by 25 points of share. A net of −10 means the reverse.

Together, the three measures place every brand in one of four positions on a Visibility–Alignment quadrant chart, with sentiment shown as point colour. In some cases, building more visibility before correcting misrepresentation makes things worse, not better. Many brands have not yet measured which position they occupy.

GBMF is a proposed open specification; validation data will be published as field data become available.

---

1. The Problem: AI Has Become a Brand Channel Without Measurement

When a buyer asks ChatGPT which enterprise analytics platforms to evaluate, or queries Perplexity for the best project management tool for remote teams, the brands returned in those answers may be perceived as authoritative. Brands absent from the response are, for practical purposes, invisible to that query. The buyer rarely goes looking for what the AI did not surface.

This is not a niche behaviour. AI-assisted discovery is growing, and the qualitative character of AI responses (synthesised recommendations rather than ranked lists of links) gives a mention a different weight than a

search result. Being mentioned in an AI response is more like being named by a knowledgeable peer than appearing on page two of a results page.

Current approaches to AI visibility measurement have three problems.

First, most tools measure clicks and impressions. These concepts do not apply directly to AI-generated responses, which typically do not produce individual page visits traceable to a specific brand mention. The measurement category is wrong.

Second, where AI-specific scores do exist, they are generally proprietary. The methodology is not published, the prompt sets are not disclosed, and scores from different vendors cannot be compared. A brand tracking AI visibility across two tools may see different numbers with no way to reconcile them or understand what either actually measures.

Third, and most important, existing tools measure at most one dimension: whether the brand is mentioned. They do not measure what is said, and they do not measure how the brand is characterised. A brand mentioned on 60% of relevant queries but consistently described with the wrong pricing, outdated product category, or pre-rebrand product name has a serious representational problem that a visibility score will never reveal. A brand mentioned accurately but consistently characterised as a poor fit, or compared unfavourably with a rival, has a different problem again.

A brand that AI systems rarely mention needs to build visibility. A brand that AI systems frequently mention but frequently misrepresent needs to address alignment first. A brand that is mentioned accurately but characterised unfavourably needs sentiment work, not more visibility. In some cases, building more visibility before fixing the alignment problem makes things worse, not better. MV, MA, and MS separate these problems so the right response can be identified.

Whether AI mention rates have measurable relationships to commercial outcomes (purchase intent, consideration, revenue) at the brand level is an open empirical question. GBMF measures representation; it does not establish that representation predicts purchase.

---

2. Three Things to Measure

Brand representation in AI has three properties. They are separable. A brand can score well on one and poorly on the others, and the combination matters as much as any score individually.

MV: how often AI mentions the brand

MV (brand-mention visibility) answers the question: how consistently does AI mention the brand across a standard set of queries relevant to its category?

The measure is built on a declared prompt set of 30 to 50 queries spanning the types of questions buyers actually ask: “what are the best tools for X?”, “compare platforms for Y”, “what do teams use when they need to do Z?”. Those prompts are run across multiple AI engines, with each prompt-engine pair queried multiple times to account for the natural variation in AI responses.

For each engine, the per-engine MV is the mean mention rate across the full prompt set. The headline MV is the equal-weighted mean of those per-engine scores. The score sits on a 0–100 scale and is interpretable as an estimated mention probability: an MV of 52 means the brand appears in an estimated 52% of AI responses to category-relevant prompts, averaged across the declared engines.

Per-engine scores are always reported alongside the headline. Engine differences are systematic rather than sampling noise, and a brand with a strong overall MV driven entirely by one engine has a concentration risk that the headline obscures.

MV is a detection measure. It operates over the full declared prompt set. It does not condition on anything: the brand was mentioned, or it was not.

MA: how accurately AI describes the brand

MA (brand-mention alignment) answers a different question: when AI mentions the brand, does it describe it correctly?

MA is measured against the Brand Profile, a structured document the brand provides. The Brand Profile specifies what is factually accurate at a given point in time: the primary product category, pricing entry point, key features, key use cases, named integrations, founding year, and any disambiguation statements the brand needs (for example, "this is not a CRM" or "this is separate from the enterprise product"). The Brand Profile is version-controlled and represents the brand's declared operational ground truth.

MA is computed only for prompts where the brand is already mentioned, and only for responses that say something evaluable about the brand. A response that includes the brand in a list without any descriptive content provides no information for alignment evaluation and is excluded. An independent evaluator scores each eligible response against the Brand Profile, field by field, and the eligible response scores within a mentioning cell are averaged to produce the cell-level value. Per-engine MA is the mean alignment score across that engine's eligible mentioning cells; the headline MA is the equal-weighted mean of the per-engine values. An MA of 67 means the average eligible mentioning cell receives an alignment score of 67/100 against the Brand Profile.

MA is a conditional measure. It conditions on the brand being mentioned. It does not condition on anything else.

MS: how AI characterises the brand

MS (brand-mention sentiment) answers a third question: when AI mentions the brand, does it speak positively or negatively about it?

MS measures sentiment directed toward the brand, not the overall tone of the response. A response that is broadly critical of a product category but singles out the brand approvingly is positive for the brand. A response that is generally favourable about the category but identifies the brand as a poor fit for certain users is negative for the brand. An evaluator instructed only to "score sentiment" will default to text polarity and mis-score exactly these cases.

Each eligible mention is classified as positive, neutral, or negative against the published codebook. MS is reported as three shares plus a net figure. The headline net is positive share minus negative share, on a -100 to $+100$ scale. The three shares are reported per engine and overall, and the headline net is the equal-weighted mean of per-engine nets. The shares-based construction matters: a net sentiment near zero may arise from uniform neutral coverage or from polarised coverage in which positive and negative characterisations both occur substantially. These are different brand situations and they call for different responses.

MS is the second conditional measure. It also conditions on the brand being mentioned. It does not condition on whether the description was accurate.

Why all three are necessary

The three measures sit at two conditioning levels. MV is a detection measure over the full prompt set. MA and MS are parallel conditional measures over the mentions MV detects. Neither MA nor MS conditions on the other: a brand can be described accurately yet positioned unfavourably, or characterised favourably yet inaccurately. All four combinations are live brand states. Improving visibility does not automatically improve accuracy. Improving accuracy does not automatically improve sentiment. The three measures separate properties that prior measurement instruments collapse or omit.

MV, MA and MS are separable measures: each addresses a distinct property, and all four combinations of accurate/inaccurate and positive/negative are live brand states. Separability is a property of the constructs. It does not assume the measures are statistically independent in field data; the same content ecosystem can drive all three, so they may correlate. The proposed validation study measures the empirical redundancy of the conditional measures with MV; the four-quadrant chart in §3 is a competitive map whose off-diagonal positions are diagnostic to the extent the measures prove non-redundant.

---

3. The Featured Diagnostic: the Visibility–Alignment Map

The featured diagnostic is the Visibility–Alignment map with MS encoded as point colour. MA is placed on the horizontal axis and MV on the vertical, because MV is the more consequential dimension (visible brands carry the larger commercial signal), so the gradient reads top-to-bottom for visibility and left-to-right for

alignment. The chart is a competitive map by default: a single category scan scores every brand from the same collected responses, so a brand's MV is informative only against where the rest of the category sits.

The four quadrants are named for position alone. These names describe where a brand sits and make no claim about sentiment:

	Low MA	High MA
High MV	Visible & Misaligned	Visible & Aligned
Low MV	Unseen & Misaligned	Aligned but Unseen

Because separability of MA from MV is a measured rather than assumed property, the off-diagonal quadrants are diagnostic to the extent the validation study finds MA and MV non-redundant.

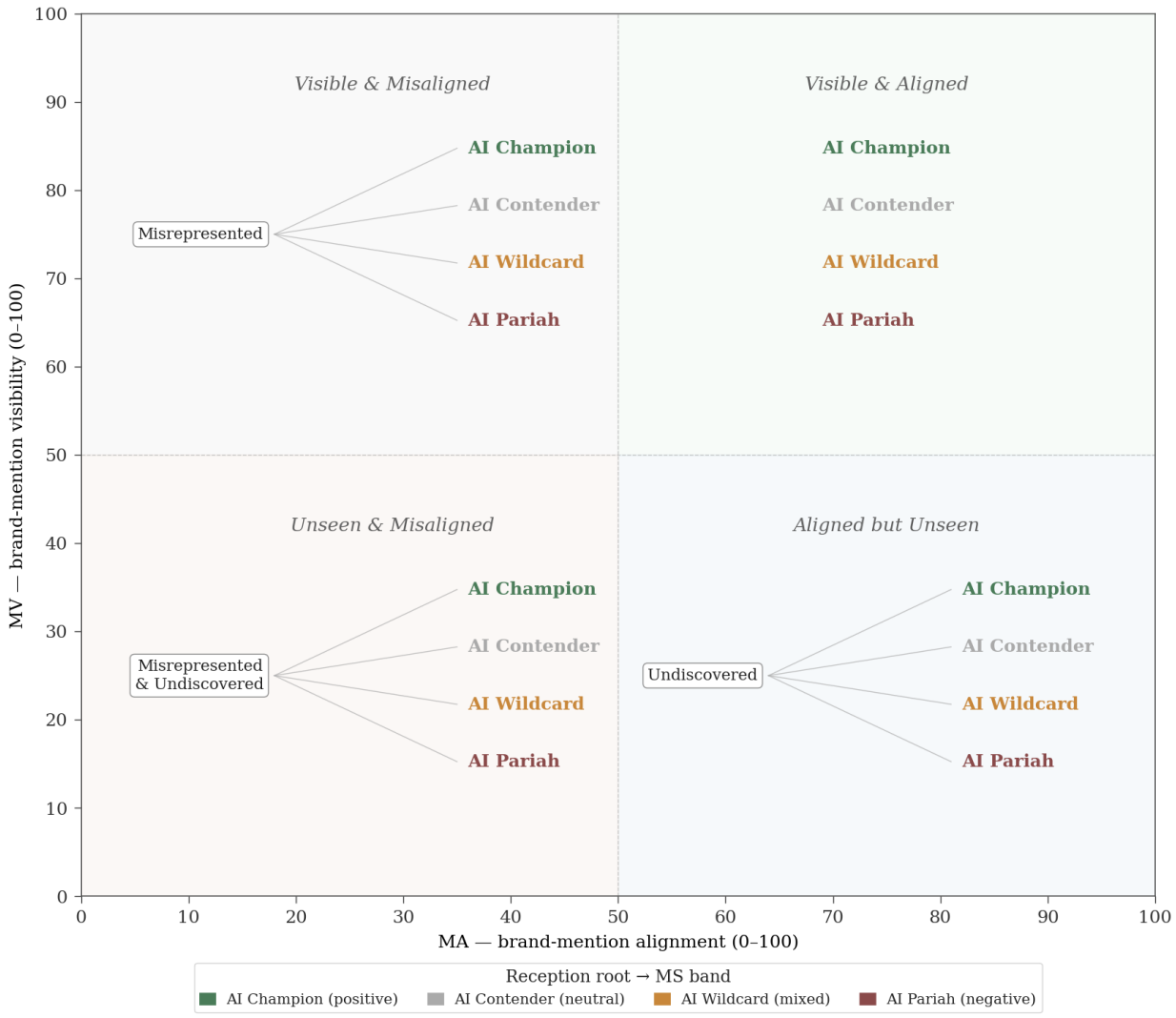
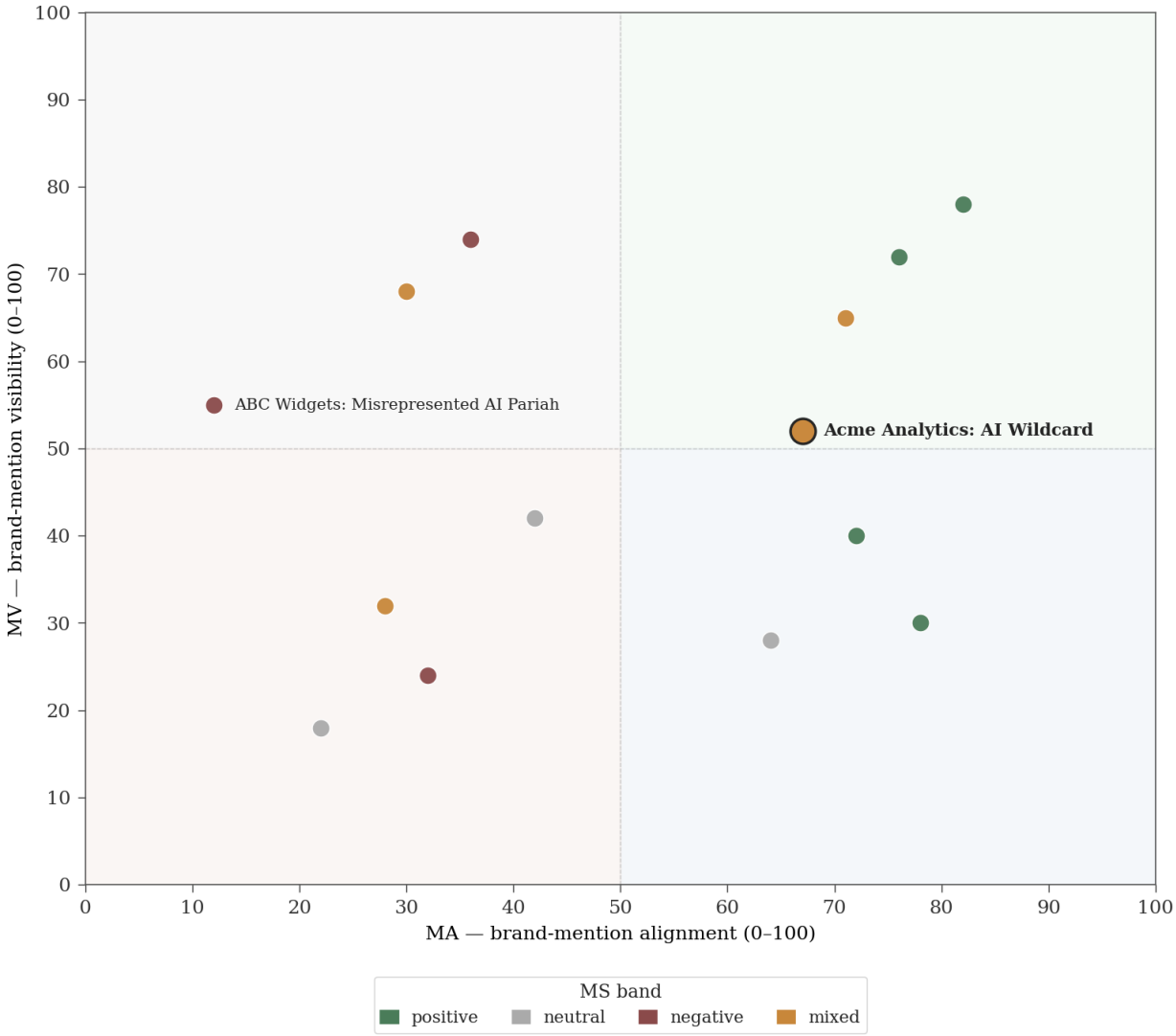


Figure 1: Teaching grid: the sixteen brand-representation states by quadrant, coloured by sentiment root

Brand-representation states. Each positional quadrant combines with the sentiment band to give a named brand-representation state. The state name has two parts. A *reception root* names where the



Example conditions (Acme): MV 52 / MA 67 / MS +25 · Engines: CLD-COP-GEM-GPT-PPX · Prompt set: B2B-ANALYTICS-v2025-11-01 · Period: 2026-06 · n=30, k=5 · evaluator: claude-opus-4-6 · eligible MA=70 (coverage 61%), eligible MS=70

Figure 2: Visibility-Alignment competitive map with Acme Analytics annotated; competitors shown as coloured dots without labels

brand stands when an answer engine assembles candidates: AI Champion for positive reception (the engine advocates for the brand), AI Contender for neutral (present in the consideration set without advocacy), AI Wildcard for mixed (on the list but dividing opinion, praised and criticised in roughly equal measure), AI Pariah for negative (characterised unfavourably). A *position prefix* names how the brand's standing departs from the visible-and-aligned ideal, with one prefix per quadrant: none when the brand is both visible and accurately described, Undiscovered when accurate but unseen, Misrepresented when visible but inaccurately described, and Misrepresented & Undiscovered when both.

Position quadrant	Prefix	Positive	Neutral	Mixed	Negative
Visible & Aligned	*(none)*	AI Champion	AI Contender	AI Wildcard	AI Pariah
Aligned but Unseen	Undiscovered	Undiscovered AI Champion	Undiscovered AI Contender	Undiscovered AI Wildcard	Undiscovered AI Pariah
Visible & Misaligned	Misrepresented	Misrepresented AI Champion	Misrepresented AI Contender	Misrepresented AI Wildcard	Misrepresented AI Pariah
Unseen & Misaligned	Misrepresented & Undiscovered	Misrepresented & Undiscovered AI Champion	Misrepresented & Undiscovered AI Contender	Misrepresented & Undiscovered AI Wildcard	Misrepresented & Undiscovered AI Pariah

The reception root states how the brand is received; the prefix states what is wrong with its standing. Neither imputes conduct by the brand: undiscovered, misrepresented, or adversely characterised describe what the answer space is doing, not what the brand has done. The reception root is determined solely by the MS display band, one root per band: Positive → AI Champion, Neutral → AI Contender, Mixed → AI Wildcard, Negative → AI Pariah. A neutral mention yields AI Contender (distinct from the positive AI Champion), and neutral is not merged into positive.

From scores to descriptor. Two lookups give the descriptor: position (MV and MA each read against the 50 boundary) sets the prefix, sentiment (the MS band) sets the reception root, and the descriptor is prefix plus root. For Acme Analytics (MV 52 visible, MA 67 aligned, shares 51/23/26 → Mixed band), the descriptor is **AI Wildcard**, a brand that is visible and accurately described and received with divided opinion.

Quadrant-level reading.

Visible & Aligned (high MV, high MA). The brand is mentioned broadly and described accurately. AI Champion is the target state. An AI Contender (neutral reception) signals an accurate, visible brand the engine carries without advocacy; the work is content that earns advocacy. An AI Wildcard signals a divided reception over an accurate baseline; an AI Pariah signals visible accurate description with negative reception. In the three off-Champion states, remediation is positioning and sentiment-driver work, not fact correction.

Visible & Misaligned (high MV, low MA). Frequent mention coupled with low accuracy. A Misrepresented AI Pariah is the commercially exposed compound state: high reach, inaccurate descriptions, adverse reception that often tracks the misrepresentation (when AI states the wrong price, the higher figure reads as "expensive" and drives the sentiment). Remediation must precede further visibility investment. The field-level MA report identifies which facts AI gets wrong; fixing the cited sources that carry the misaligned claims is the primary action, and the negative sentiment frequently resolves with the facts. A Misrepresented AI Wildcard sits in the same quadrant with polarised reception over the inaccurate descriptions; the same alignment-first sequence applies.

Aligned but Unseen (low MV, high MA). The brand is mentioned infrequently but accurately described when it is. The accuracy foundation is in place; the primary problem is discoverability. An Undiscovered AI Champion can scale visibility safely; an Undiscovered AI Contender (neutral reception) likewise; the visibility push needs no sentiment work first. An Undiscovered AI Wildcard or Undiscovered AI Pariah signals that the

brand's few mentions carry polarised or adverse reception, so sentiment-driver analysis precedes amplification, since scaling visibility without first resolving the reception drivers amplifies the negative coverage.

Unseen & Misaligned (low MV, low MA). Minimal AI presence, and when the brand does appear the descriptions are not reliably accurate. Both dimensions need work, and the order matters: attempting to build visibility before establishing an accurate baseline risks arriving at the Visible & Misaligned position at higher scale. The sensible starting point is specifying the Brand Profile precisely and building visibility from that foundation. Where the brand is also below the eligible-cell floor on the conditional measures, the response is data first: build sufficient mention volume that MA and MS become reportable before sequencing remediation against them.

Sentiment overlay and Mixed.

MS sits as point colour on the chart: positive (green), neutral (grey), mixed (amber), negative (red). Neutral and mixed are distinct bands and are not merged: neutral indicates the absence of directional characterisation; mixed indicates co-present praise and criticism (both shares $\geq 25\%$). A high MV with mixed sentiment carries different commercial risk than a high MV with uniformly positive sentiment, even though both might net to the same number; the mixed band is the diagnostic that net alone hides.

In some categories, sentiment work matters more than alignment work; in others, less. The framework imposes no single priority order; its discipline is seeing all three properties separately, so the order can be chosen on evidence.

Insufficient data and AI Absent. A brand that clears the eligible-cell floor (10 eligible mentioning cells per engine, see §5) on both conditional measures resolves to one of the sixteen brand-representation states. A brand below the floor on alignment, on sentiment, or on both carries insufficient data for the affected measure. It takes no state label, since the position requires a readable MA and the reception root a readable MS, and is reported by its readable metrics with the shortfall flagged, in any quadrant. The shortfall is per-measure: a brand may have readable MA but insufficient MS, or the reverse, or both insufficient. A brand with MV = 0 produces no mentioning cells. MA and MS are undefined, and the brand is reported as **AI Absent**. AI Absent is not a reception root, carries no prefix, and sits off the teaching grid.

The boundaries at MV = 50 and MA = 50 are presentation conventions for the figure; they are declared as such in figure notes and do not enter any computation. The underlying scores are continuous.

---

4. How MV Is Measured

GBMF is a proposed specification. The measurement procedures in §§4–6 describe the intended methodology; calibration thresholds are stated as provisional conventions pending empirical calibration.

MV measurement begins with a standardised prompt set. For each category, 30 to 50 prompts are generated across four query types: categorical queries ("what are the leading platforms for X"), comparative queries ("compare options for Y"), use-case queries ("what do teams typically use when they need to do Z"), and problem-solution queries ("what's the best solution for companies struggling with W"). The prompt set spans the full range of ways buyers approach a category and does not contain any brand name from the measured competitive set, so a single scan scores every brand in the category from the same stimulus. The generation process and governance rules are in the technical paper.

These prompts are run across the declared engine set. The current standard engine list comprises ChatGPT, Claude, Microsoft Copilot, Gemini, AI Overviews, and Perplexity, identified by three-character codes (GPT, CLD, COP, GEM, AIO, PPX). Implementations may declare a different engine set; the declared set is part of every published score. Each prompt is run a minimum of three times per engine to account for the natural variation in AI responses. Because AI engines are probabilistic, the same prompt can produce different outputs across runs. A brand mentioned on one run but not the next has a different visibility profile than a brand mentioned reliably on every run. Measurement that does not account for this variation produces misleading results.

For each prompt and each engine, the per-engine mention rate is the fraction of runs in which the engine mentioned the brand. Per-engine MV is the mean of those rates across the full prompt set, expressed on

a 0–100 scale. Headline MV is the equal-weighted mean of the per-engine values. This per-engine-first aggregation is used throughout the framework: it makes the headline reconcile exactly with its parts, and it keeps the score interpretation longitudinally stable. A score should change because AI mention behaviour changed, not because market shares among the declared engines shifted.

For published benchmark figures, the collection requirement is $n \times k \geq 150$ observations per engine. At $n = 30$ prompts and $k = 5$ runs, this threshold is met exactly. The $k \geq 3$ minimum applies to all measurement, including monitoring scans; only scans feeding published benchmark figures must clear $n \times k \geq 150$.

Three companion statistics are reported alongside the headline MV, never merged into it:

- **Reach:** the percentage of prompts on which the brand is mentioned at least once. An MV of 40 achieved by moderate presence across many prompts differs from one achieved by high presence on a narrow slice.
- **Intensity:** the mean mention rate among prompts where the brand was mentioned. Together with reach, intensity describes the shape of the mention distribution.
- **Concentration:** the Herfindahl-Hirschman Index over the prompt-level contribution shares. Uniform baseline is 1/n: a multiple of uniform is the readable quantity. High concentration with a given MV signals fragile, narrow visibility.

For Acme Analytics, the worked example brand in the technical paper, MV = 52 with reach 87%, headline intensity 0.68 (per-engine range 0.57–0.73), and HHI 0.047 (approximately 1.4× uniform), indicating broadly distributed visibility close to the uniform baseline.

---

5. How MA Is Measured

MA measurement starts with the Brand Profile. This is a structured document the brand creates, with support from the GBMF specification's published template, specifying what is factually correct about the brand at a given point in time. It covers the primary product category, the pricing entry point, key features, key use cases, named integrations, founding year, company type, and any important disambiguation statements.

The Brand Profile distinguishes between types of claims through a three-tier model:

Tier	Nature	Examples	MA treatment
1	Externally verifiable facts	Founding year, company type, product category, pricing	Full weight
2	Brand-stated facts (operationally authoritative)	Feature descriptions, use cases, named integrations	Full weight
3	Positioning claims	"Best for enterprise", "leading platform"	Excluded

A Brand Profile used for benchmark MA publication must contain at least three Tier 1 fields. This is a normative requirement of the specification, not a governance aside: Tier 1 fields are the only externally auditable component of MA's ground truth, and the Tier 1 minimum is the framework's convergent-validity floor. Tier 3 positioning claims are excluded from scoring entirely; whether AI agrees with competitive positioning is not a factual alignment question.

MA is computed only for prompts where the brand is mentioned and only for responses that say something evaluable. A response is **eligible** for MA if it triggers at least two evaluable Brand Profile fields (fields where the response makes claims that can be evaluated against the declared value). Bare-name mentions in lists, with no descriptive content, are excluded. This eligibility rule ensures near-zero-information observations do not distort the score.

Eligible responses are scored field by field by an independent evaluator: an LLM from a different model family than the engine being evaluated, ensemble-averaged across three independent evaluations per response, and calibrated against a human gold standard before any score is published. Per-engine MA is the mean alignment score across that engine's eligible mentioning cells. Headline MA is the equal-weighted mean of the per-engine values, including only engines that clear the eligible-cell floor of 10. The cross-family evaluator selection guards against shared pre-training bias, where an evaluator from the same model family may reproduce identical training-data errors as the engine being evaluated and artificially inflate the score.

MA coverage (eligible mentioning cells as a proportion of total mentioning cells) is reported alongside every MA score. Coverage is a first-order diagnostic, not a footnote. If most mentions are bare-name list inclusions, the brand is named but never described in AI responses. This is a distinct representational state with its own remediation path: the brand needs describable, citable claims in circulation. Coverage below 50% is treated as a caution condition.

A few brand situations are particularly exposed to low MA. Recently rebranded or repriced brands are at the highest risk, because AI systems trained before the change continue to describe the old product and the web still contains more content about the old version than the new one. Category ambiguity creates a related risk. When third-party coverage frames a brand differently than the brand frames itself, AI systems follow the dominant external framing, a problem that intensifies when a brand competes closely with a better-known rival, where AI descriptions sometimes absorb competitor characteristics entirely.

The Brand Profile must be version-controlled. Any update requires a new version identifier and creates a marker on the MA series flagging that the reference standard has moved. The MA series remains continuous through the marker; the marker is the practitioner's instruction that cross-boundary MA values are measured against different declared facts and are not directly comparable as a performance change. MV does not condition on the Brand Profile and runs unbroken through the event.

For Acme Analytics, MA = 67 with coverage 61%. Pricing is the dominant field-level misalignment: approximately half of all eligible mentioning cells included a pricing claim outside the ±15% tolerance. Category classification and feature claims are well-aligned. This field-level breakdown is the input to remediation planning.

---

6. How MS Is Measured

MS measurement applies the same eligibility discipline as MA, with a different definition of "eligible". A response is eligible for MS if it contains substantive evaluative characterisation of the brand beyond bare-name inclusion. A response listing the brand in a set ("top tools include X, Y, Acme, and Z") with no evaluative content is not eligible for MS. The MS eligibility criterion may yield a different count than the MA criterion: a response may be MA-eligible (triggers at least two fields) but MS-excluded (no evaluative characterisation), or MS-eligible (evaluative characterisation present) but MA-excluded (fewer than two fields triggered).

Eligible responses are classified as positive, neutral, or negative by the evaluator using the published codebook. The evaluator instruction is explicit: classify the sentiment expressed toward the brand, not the overall text tone. A response that criticises the product category but praises the brand is positive for the brand. A response that is generally favourable about the category but identifies the brand as a poor fit for certain users is negative. An evaluator scoring text polarity instead of brand-directed sentiment will systematically mis-score these cases, producing non-comparable results across implementations that differ on this point.

Per-engine shares (% positive, % neutral, % negative) are computed first. Per-engine net is positive share minus negative share, on a -100 to $+100$ scale. The headline shares and headline net are the equal-weighted mean of per-engine values over engines clearing the eligible-cell floor. This per-engine-first aggregation mirrors MV and MA.

The shares-plus-net reporting structure is preferred over a single averaged bipolar score because it distinguishes two conditions that an average collapses: a high neutral share (most descriptions are neither favourable nor unfavourable) and high positive plus high negative shares netting near zero (the brand is

praised and criticised in roughly equal measure). Both may produce a net near zero, but they represent different brand situations and different remediation needs. The shares-and-net structure makes the difference visible at a glance.

For Acme Analytics, the headline shares are 51% positive / 23% neutral / 26% negative, headline net +25, in the Mixed band. The brand is positively characterised on integration breadth and ease of use; the negative characterisations are concentrated on pricing, driven by AI responses stating the wrong price, which then read as "expensive relative to alternatives". Because the negative sentiment tracks the pricing misalignment identified in MA, fixing the price facts in cited third-party sources is likely to improve both MA and MS simultaneously.

LLM evaluators exhibit a documented positivity lean on annotation tasks. The human calibration step tests for a systematic positivity offset on brand-directed sentiment specifically and corrects it before production use. MS is the framework's most interpretive measure: sentiment classification is less determinate than factual alignment checking, and the expected reliability ceiling for MS sits below MA's. Both ceilings are reported and distinguished.

---

7. Tracking Over Time: Governance and Version Control

A single MV, MA, or MS measurement tells the brand where it stands today. A dated series of measurements tells the brand whether its position is changing, whether actions are working, and whether AI systems are drifting in how they describe it. Time-series data is where measurement becomes strategically useful.

The temporal view is the dated series of unmodified scores. No score is ever adjusted. Changes are marked, decomposed, and explained. Four event types are handled under one no-adjustment principle.

Engine model updates (world changes). A model version update behind a declared engine surface flows through the series as real movement. The engine is defined as the user-facing surface, not the model version behind it. An update to the underlying model is a real change in the brand's answer-space representation; a score change at a known model update is the detection signal, analogous to a traffic pattern change at a search algorithm update. Any mechanism that adjusts scores at model update boundaries launders this signal.

Prompt-set version transitions. When the prompt set is updated (updates are permitted only on documented category-change triggers), both versions run concurrently for one overlap scan window. At the version boundary, the series carries a version marker and publishes three quantities: the old-set score, the new-set score, and a decomposition stating how much of the difference is movement on familiar questions versus position in the new question space. An example: "v1 held at 54; v2 opens at 46; the gap is absence from the new capability queries added to the category in this cycle." The series continues with new-set scores. The old-set history stands as measured; no splicing or rescaling occurs.

Evaluator changes. Any evaluator model change requires full re-validation against the human gold standard before new scores are published. The evaluator version is documented in every reporting block. Drift detection runs continuously through periodic re-evaluation of a small human-coded sample.

Brand Profile version changes. A Brand Profile update creates a marker on the MA series only. The MA series continues through the marker as a single series; what it carries at the boundary is a caution, not a severance. MA either side of the marker is measured against different declared facts, so the cross-boundary gap is not directly interpretable as a performance change. The practitioner reads movement across the marker knowing the reference moved. The MV series does not condition on the Brand Profile and runs unbroken through the event.

The marker on a Brand Profile change is stronger than the marker on a prompt-set transition, and the reason matters. At a prompt-set transition the construct stays constant (same definition of visibility and alignment, different question space), so the cross-boundary gap is interpretable and decomposable. At a Brand Profile change the reference standard for "correct" has itself moved, so the cross-boundary MA gap is not directly interpretable as performance. The framework distinguishes the two because they call for different practitioner readings.

The practical rhythm for most brands is quarterly measurement for tracking, with a deeper diagnostic audit at significant business events: a rebrand, a repricing, a product launch, an acquisition. Those events are precisely when MA and MS are most likely to deteriorate and least likely to be monitored.

Measurement cadence may also vary by engine type. Engines that use live web retrieval (Perplexity, ChatGPT Search, AI Overviews) can shift more rapidly as indexed content changes. Pure generative engines shift more slowly, typically on the cycle of model updates. A brand with a mixed-engine profile may need higher-cadence monitoring than a brand whose visibility is concentrated in engines of one type.

Every published score carries a mandatory reporting conditions block: engine list, prompt set version, measurement period, n and k, evaluator model and version, eligible cell counts, and coverage statistics. A score without its conditions cannot be compared to another score. Conditions are part of the result.

---

8. What This Enables

The most immediate value of GBMF is diagnostic: knowing where the brand sits on the Visibility–Alignment plane, what sentiment band it occupies, and what is driving its scores. The framework also enables several things that were not previously possible.

Prioritising investment correctly. A brand team that knows only that its AI presence is weak cannot determine the right response. MV, MA, and MS together make the priority order explicit. Low MV with high MA warrants a content investment to expand mention reach. High MV with low MA warrants accuracy remediation before further visibility work. High MV and high MA with adverse sentiment in the Mixed or Negative band warrants sentiment-driver analysis: which characterisations recur, and which sources carry them. Acting on the wrong diagnosis wastes the investment.

Evaluating whether actions work. Structured data improvements, content updates, PR campaigns, and entity optimisation all claim to influence AI visibility. With a standardised measurement instrument and a version-controlled time series, brands can track whether those actions actually change MV, MA, or MS, on which prompts and engines the changes show up, and over what cadence. Without this, AI marketing investment is largely untestable.

Cross-brand and category benchmarking. Because GBMF is produced by a published, open methodology with standardised prompt sets, a normative Brand Profile structure, and a published scoring rubric, scores are comparable within a governed measurement design, and cross-category comparisons are possible when the prompt sets, engine sets, and reporting conditions are declared. Within a category, every brand is scored from the same category benchmark prompt set, so the Visibility–Alignment quadrant is a directly comparable competitive map by default. Cross-category readings remain interpretable, but they involve different prompt sets and therefore different measurement environments. This kind of structured comparability is not possible with proprietary scores from different vendors using different, undisclosed methods.

Due diligence and brand audits. GBMF can provide structured AI brand exposure data for acquisition due diligence and brand audits. A brand with high MV and low MA carries a representational risk that traditional brand health metrics do not surface; a brand with minimal presence in a category that increasingly routes discovery through AI has an exposure worth documenting; a brand with adverse sentiment alongside healthy visibility and alignment has a category-positioning risk that the headline scores alone obscure. This application is conditional on the framework being empirically calibrated; it is a projected use case, not a current product claim.

Research and competitive intelligence. Because the methodology is open, it can be applied to competitors as well as to the brand itself. Understanding where competitors sit on the Visibility–Alignment plane, and what sentiment bands they occupy, provides a cleaner competitive picture than share-of-voice metrics, which do not separate mention from accurate description or favourable from unfavourable characterisation.

The open specification is a deliberate choice. Proprietary scores with undisclosed methodologies cannot be independently validated, compared across vendors, or built upon. GBMF is published as an open specification so that anyone implementing it using the same prompt set, Brand Profile, evaluator architecture, and scoring rubric produces comparable scores. Aiviara Research maintains a reference implementation; independent implementations are encouraged.

9. Getting Started

Implementing GBMF measurement requires four inputs: a Brand Profile, a declared prompt set, access to the declared AI engines, and a validated evaluator.

The Brand Profile takes a few hours to assemble for the first time. The specification's published template covers each field category. The fields that require the most care are the disambiguation statements (the things AI systems commonly get wrong about the brand that need explicit correction) and the decision about which fields belong in Tier 1 (externally verifiable) versus Tier 2 (brand-stated operational). A benchmark-grade Brand Profile must contain at least three Tier 1 fields. Getting these right matters because the Brand Profile is what MA is measured against; a sparse Brand Profile produces low or unreliable MA, not high MA.

The category benchmark prompt set is generated under the rules in Appendix A of the technical paper and assigned a version identifier in the format `[CategoryID]-v[YYYY-MM-DD]`. The set is held immutable across measurement periods within a version, and the same set scores every brand in the category. Brand-diagnostic prompt sets (which may name the focal brand) are maintained separately under a `[BrandID]-v[YYYY-MM-DD]` namespace; their scores are non-comparable to MV/MA/MS benchmark figures.

Engine access requires running prompts at the cadence and volume needed for $n \times k \geq 150$ observations per engine, with automation, credential management, and rate-limit handling across multiple engines. Implementation details are covered in the technical paper. Practitioners must review the terms of service of each commercial engine before conducting systematic automated querying at collection scale.

The evaluator must be an LLM from a different model family than the engines being evaluated, with a documented pinned model version. Before any score is published, the evaluator clears validation targets against a human gold standard: Cohen's kappa or Krippendorff's alpha ≥ 0.75, LLM-human score correlation $r \geq 0.80$, and for MS specifically a positivity-offset test. Validation is repeated on every evaluator model change. The framework also specifies a determinism-tested control arm: a locally hosted, version-pinned open-weight model whose behaviour is stable by construction and which serves as the noise-model validator for the measurement pipeline.

The full mathematical specification, scoring rubric, sentiment codebook, evaluator prompts, and formal design rationale are in the companion technical paper: *Brand Representation in AI Answers: An Open Specification for Measuring Visibility, Alignment and Sentiment.* The full per-prompt per-engine mention rates for the Acme Analytics worked example are published as supplementary data so the headline figures reproduce from an open dataset.

For practitioners new to the formulas, a companion document, *The Generative Brand Mention Framework: A Formula-by-Formula Primer*, walks through every formula with worked examples and Python implementations.

Further Reading

Khan, B. (2026). *Brand Representation in AI Answers: An Open Specification for Measuring Visibility, Alignment and Sentiment.* Aiviara Research. Working paper, June 2026.

Aggarwal, P., et al. (2024). GEO: Generative Engine Optimization. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* https://doi.org/10.1145/3637528.3671900

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. https://aclanthology.org/2024.eacl-demo.16/

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12076–12100. https://aclanthology.org/2023.emnlp-main.741/

Schulte, J., Bleeker, M., & Kaufmann, P. (2026). Don't measure once: Measuring visibility in AI search (GEO). arXiv:2604.07585 [cs.IR].

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Sage.

Aiviara Research is a research and publishing initiative. For more information visit aiviara.com/research.