Brand Representation in AI Answers: An Open Specification for Measuring Visibility, Alignment and Sentiment

Badrudin Khan

Aiviara Research

Working paper · Methodological specification · June 2026

Abstract

This paper specifies the Generative Brand Mention Framework (GBMF), a proposed open measurement specification for brand representation in AI-generated responses. GBMF comprises three separable measures. Brand-mention visibility (MV) estimates the probability that the brand appears in an AI response to a category-relevant prompt. Brand-mention alignment (MA) measures, given a mention, the consistency of the description against the brand's own declared facts, not external truth. The "alignment" construct here concerns factual correspondence, distinct from the AI-alignment construct that concerns model behaviour. Brand-mention sentiment (MS) measures, given a mention, whether the response speaks positively or negatively about the brand. MA and MS each condition on mention.

The three measures are necessary because the underlying mechanisms are separable. A brand can be mentioned yet described inaccurately, or described accurately yet positioned unfavourably, and these are different remediation problems.

MV and MA are reported on a 0–100 scale; MS as three percentage shares plus a net figure on −100 to +100, over a governed prompt set and a declared engine set.

GBMF specifies formal definitions, error quantification, companion statistics, evaluator architecture with a determinism-tested control arm, governance requirements, and pre-stated falsifiability criteria. An end-to-end worked example accompanies the specification, with per-prompt per-engine mention rates published as supplementary data. The framework is published in full to enable open, governed, falsifiable measurement.

Keywords: Generative Brand Mention Framework, GBMF, MV, MA, MS, AI visibility, AI search, answer engine, answer engine optimisation, AEO, GEO, brand monitoring, brand visibility, brand sentiment, generative AI, large language models, measurement methodology

Validation status. This paper specifies MV, MA, and MS as proposed open measurement instruments. It does not report empirical validation results. Prospective validation targets are stated in §9.6. Calibration conventions (the $n \times k \geq 150$ collection rule, the eligible-cell floor of 10) are stated as provisional conventions, pending empirical calibration against observed variance and score dispersion. A validation study along the lines set out in §9.6 is intended, and if conducted, results would be released as field data become available.

AI-assisted preparation. This manuscript was prepared with the support of AI drafting and critique tools. The author reviewed, edited, and verified the substance of every section, including all claims, formulae, references, and conclusions, and accepts full responsibility for them. AI tools were not used as the source of theoretical commitments, empirical findings, or methodological positions. Full disclosure appears at the end of the paper.

1. Introduction

AI search engines are changing how brands are discovered and evaluated. When a buyer asks ChatGPT for "the best project management tools for remote teams", or queries Perplexity for "enterprise data warehousing solutions", the brands mentioned may be perceived as authoritative. Absent brands are, for those queries, invisible.

A brand's position in this answer space turns on three separable conditions that prior measurement instruments collapse. Whether the brand appears at all is the foundational question. A brand can invest in content and structured data and still barely register. Whether it is described correctly is a distinct question, since a brand mentioned in 40% of relevant responses still faces a real problem if most of those mentions state the wrong price or misidentify the product category. And whether the response speaks positively or negatively about it is a third, separable question, because a brand can be accurately described but consistently positioned unfavourably against competitors.

Category-level AI prompts can be understood as a machine analogue of category-cued, unaided brand recall, where the prompt supplies the category cue and the response reveals which brands are available in the answer space (Keller, 1993; Lavidge & Steiner, 1961). The respondent is the AI. The test items are the prompts. The brand's score reflects how often it appears in the answers, and what those answers say about it. This framing motivates which properties are measured: presence, correctness, sentiment.

Measurement of brand presence in AI-generated answers is an active and growing area, spanning practitioner benchmarks, academic studies of source citation, and proprietary commercial scoring. This paper specifies the Generative Brand Mention Framework (GBMF), an open framework for the three separable properties of brand representation in AI answers: brand-mention visibility (MV), brand-mention alignment given mention (MA), and brand-mention sentiment given mention (MS). MA and MS each condition on mention. All three are specified under declared prompt sets, declared engine sets, declared evaluator architecture, and pre-stated conditions for comparability, held fixed and published in full so that scores are comparable across studies and over time.

1.1 Contributions

The paper's central contribution is the Generative Brand Mention Framework (GBMF), an open measurement specification that integrates three separable measures with declared prompt sets and engine sets, cell-first denominators, evaluator architecture, and pre-stated falsifiability thresholds. The per-measure formulas are deliberately standard, a mean mention probability for MV, a mean alignment score for MA, and share-based net sentiment for MS, formally defined with error quantification and companion statistics in §§4–6.

Governance is specified at the level the framework needs to be open: prompt-set construction (Appendix A), Brand Profile version control with a Tier 1 minimum (§5.2), evaluator architecture including a determinism-tested control arm (§7.5), and a single no-adjustment principle covering engine updates, prompt-set versions, and Brand Profile changes (§8.4).

The proposed validation programme states in advance the specific empirical thresholds at which the framework would be revised, alongside a concurrent expert-panel criterion that would anchor MA's convergent validity outside self-report (§9.6).

2. Related Work

2.1 GEO and AEO research

Aggarwal et al. (2024) presented a controlled experiment modifying 10,000 passages against leading LLMs, finding citation additions increased AI visibility by up to 40% while keyword-based optimisation had negligible or negative effects. That work introduced two formally defined impression metrics measuring source-level visibility.

Chen et al. (2025) examined AI search mention behaviour across multiple engines and found systematic bias toward earned media over brand-owned and social content, with significant cross-engine differences in domain diversity, freshness, and prompt-phrasing sensitivity. Schulte, Bleeker, and Kaufmann (2026) demonstrated that AI search responses vary probabilistically across runs and over time, and that single-observation measurements are unreliable. Kumar et al. (2025) empirically tested the GEO-16 framework, identifying Metadata/Freshness, Semantic HTML, and Structured Data as the strongest predictors of AI mention across 1,702 citations.

Practitioner research has produced directionally useful benchmarks that cannot be synthesised due to methodological heterogeneity. AI assistants cite content averaging 1,064 days old versus 1,432 days for organic results (Ahrefs, 2025a), and favour pages with high Domain Rating (Ahrefs, 2025b). AI referral traffic represents about 1.08% of total website traffic (Conductor, 2026, data collected May–October 2025). OtterlyAI (2026) found no measurable positive effect of llms.txt on AI visibility.

Several practitioner sources cited here (OtterlyAI, Ahrefs, Conductor) are tools with commercial interests adjacent to those of the author. They are cited as publicly available practitioner evidence with commercial-interest limitations stated, and their findings are attributed transparently.

2.2 Information quality and faithfulness evaluation

Measurement frameworks for AI-generated text quality have a natural precedent in the data quality literature. Wang and Strong (1996) proposed a foundational taxonomy of data quality, identifying accuracy as an intrinsic property of high-quality data within a framework of four quality categories (Intrinsic, Contextual, Representational, and Accessibility). Their definition of accuracy, the extent to which data values conform to their true values, provides the conceptual basis for MA. In the AI context, the data values are the claims made by AI engines about a brand, and the true values are specified in the brand's authoritative profile.

FactScore (Min et al., 2023) and SAFE (Wei et al., 2024) are the closer prior art for MA and the primary comparison. Both decompose long-form LLM outputs into atomic claims and evaluate each against an external knowledge source. FactScore uses a fixed reference corpus (Wikipedia biographies). SAFE issues live web searches per claim and adjudicates via an LLM judge. Each measures factual precision against the world as captured in an external corpus. MA measures something deliberately different, namely representational consistency between AI-generated brand descriptions and the brand's own version-controlled declared facts. The estimand is different because the operational question is different. A brand owner asks whether the answer space carries the brand as the brand describes itself, not whether each atomic claim is recoverable from an external corpus.

The brand-controlled choice is contested. It is what makes the measure auditable across measurement periods (the Brand Profile is versioned and held fixed, while an external corpus is not), and what allows MA to score correctly on facts no external source has caught up to, such as a price change, a renamed product, or an updated category positioning. It is also what allows a brand, in principle, to curate a profile that simply matches what AI already says. The Tier 1 minimum (§5.2) is the structural constraint against that case, requiring at least three externally verifiable fields in any Brand Profile used for benchmark MA publication. Tier 1 is a floor on external anchoring, not a substitute for the corpus-grounded approach FactScore and SAFE use, and the full convergent-validity argument is in §10.2.

RAGAS (Es et al., 2024) is the more distant adjacent. It measures faithfulness as the fraction of claims in a generated answer that can be inferred from the retrieved source context for retrieval-augmented generation. MA applies a cognate evaluation mechanism in a different direction. Where RAGAS asks whether a response is faithful to retrieved documents, MA asks whether AI-generated brand descriptions are faithful to the brand's authoritative profile. The structural similarity ends at the evaluation mechanism, and MA is not an extension of RAGAS.

2.3 Sentiment measurement

For brand-directed sentiment in AI-generated content, the framework's methodological lineage runs through media monitoring and social-listening practice, which employs tonality coding against published codebooks with inter-rater reliability reporting (Krippendorff, 2004). The shares-plus-net reporting structure used for

MS is established in polling favourability measurement (net approval) and in the commercial origin of the net-favourability form (net promoter; Reichheld, 2003). MS uses Reichheld as the reporting-structure analogue only, not as methodological authority on the choice between shares-plus-net and averaged bipolar scales. The methodological weight for the construct sits with the stance-detection and content-analysis literatures cited in §6 and Appendix D.

LLM-as-annotator validation is an active area. Recent computational social science work establishes that LLMs can achieve inter-annotator agreement comparable to human crowdworkers on classification tasks, with human calibration as the reliability ceiling (Gilardi et al., 2023; Wang et al., 2023). MS relies on this literature for its production architecture.

---

3. Background

3.1 Measurement background

LLMs mention brands through two distinct mechanisms.

Training-data mention occurs when a brand has sufficient representation in the model's pre-training corpus. Mention rates reflect third-party coverage indexed before the training cutoff. Chen et al. (2025) find that AI search shows systematic preference for earned media over brand-owned content. This mechanism also introduces the alignment problem. Training data may embed descriptions that were aligned with the brand's profile at time of indexing but have since become outdated, or may carry errors from unreliable sources. For training-data-dominant engines, MA misalignments may persist until the next training cycle.

Real-time retrieval mention occurs when the model performs a live web search, as with ChatGPT Search and Perplexity. Mention rates relate to search engine rank, though editorial authority is a stronger predictor than position alone among pages with trackable organic rankings (Ahrefs, 2025b). Real-time retrieval reduces but does not eliminate alignment risk. Cited sources may contain errors, outdated pricing, or competitor-generated content.

These distinctions affect interpretation of all three metrics: scores differ by engine type, and that variation is information rather than noise. Per-engine decomposition makes it visible.

Mention rates are also undefined without specifying the prompts used to elicit them. The prompt set defines the answer space being measured, not just the sample drawn from it. A set covering only comparative queries produces a systematically different MV from one that includes problem-solution and use-case queries, so two studies with different prompt sets are measuring different things even when both report a number labelled MV. Schulte et al. (2026) show that single-observation measurements are unreliable under probabilistic response variance. The framework therefore requires standardised prompt sets with version control and repeated measurement runs per prompt-engine pair, with the prompt-set version identifier carried in every published score condition.

3.3 The Generative Brand Mention Framework

The two mechanisms above generate three measurable quantities for any brand in a defined answer space. MV is a detection measure over the full declared prompt set. MA and MS are parallel conditional measures over the mentions MV detects. Both condition on the same event (mention), and neither conditions on the other. All four combinations of accurate/inaccurate and positive/negative are live brand states, so a brand can be accurately described yet positioned unfavourably, or described favourably yet inaccurately.

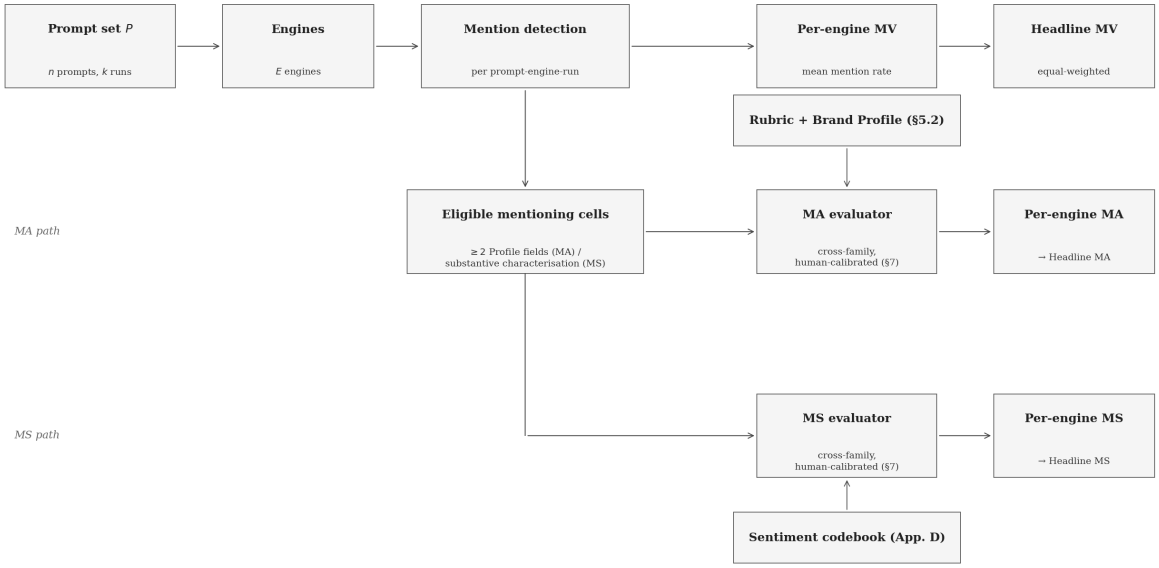


Figure 1: The two-level measurement pipeline.

Mention detection produces per-engine MV. The mention stream forks into two parallel evaluators for MA and MS. Evaluator boxes annotated "cross-family, human-calibrated (§7)".

---

4. MV: brand-mention visibility

4.1 Estimand

MV estimates the probability that brand B is mentioned by name in an AI response to a category-relevant prompt, with respect to a declared engine set, prompt set version, and measurement period.

The engine set, prompt set version, and measurement period are part of the estimand's definition. An MV score pertains to a single scan window. Change over time is represented as a series of dated scores, never as a blend of scans (see §8).

A **brand mention** is the explicit naming of a brand entity within an AI-generated response, regardless of whether a source URL is attached. A brand mention is distinct from a source citation (a cited URL or document). A response may mention a brand without linking to a source, and a source may be cited without the brand name appearing in the response text. MV measures brand mentions only.

An **engine** is a distinct end-user AI response system with an independently operated retrieval or generation stack. Multiple interfaces backed by the same underlying model family are treated as separate engines when they exhibit materially different retrieval behaviour or response construction. The engine set must be declared and held constant across all measurement periods for a given brand.

A **prompt set** P is a standardised, version-controlled set of n prompts associated with a competitive category. Two prompt-set types are maintained: a category benchmark set (unbranded, used for cross-brand MV/MA/MS comparison) and brand diagnostic sets (may contain the focal brand, used for single-brand reporting only). Both are generated using the methodology in Appendix A, assigned a version identifier, and held immutable across measurement periods. Unless otherwise stated, "prompt set" refers to the canonical category benchmark set.

The canonical benchmark prompt set is category-level and does not contain the focal brand name. Branded comparative prompts (those naming the focal brand or a specific competitor) are confined to diagnostic, single-brand reporting and are not used for cross-brand MV benchmarking. The distinction prevents mechanical inflation of MV from prompts that force inclusion of the focal brand.

4.1.1 Unit of analysis and measurability

The framework has two measurement tiers. MV is unconditional. It is computed over the full declared prompt set, so every prompt-engine-run contributes whether or not the brand is mentioned, and MV is estimated on the full n × k observations regardless of how often the brand appears. MA and MS are conditional on mention. Each is computed only over the prompt-engine cells in which the brand is mentioned. A prompt-engine cell is the set of k repeated runs for one prompt on one engine.

For MA and MS, scores are aggregated cell-first. Eligible scores are averaged within each mentioning cell, and the per-engine value is the mean of those cell-level scores across eligible cells. Repeated mention runs within a cell affect the stability and confidence of that cell's value, not its weight in the per-engine figure. A prompt-engine condition that mentions the brand on all k runs and one that mentions it on a single run each contribute one cell. The denominator for MA and MS is therefore eligible mentioning cells, not individual responses.

Because MA and MS share the conditioning event and the cell denominator, they share a sample fate. A brand with few mentions has few eligible cells for both, and both are governed by the same eligible-cell floor (§5.3). Below that floor a brand carries insufficient MA or MS data on the affected conditional measure, in any quadrant of the featured diagnostic. The shortfall is per-measure: a brand may have readable MA but insufficient MS, or the reverse, or both insufficient. MV remains reliable for such a brand because MV is unconditional, and a low MV is itself a sound measurement that the brand rarely appears. A brand with MV = 0 produces no mentioning cells. MA and MS are undefined, and the brand is reported as AI Absent.

4.2 Definitions

Let $c(p, e, s) \in \{0, 1\}$ be the mention outcome for prompt p, engine e, and run s, with $p \in \{1, \ldots, n\}$, $e \in \{1, \ldots, E\}$, and $s \in \{1, \ldots, k\}$; n is the number of prompts, E the number of engines, and k the number of runs per prompt-engine pair.

The per-engine mention rate per prompt:

$$r_e(p) = \frac{1}{k} \sum_{s=1}^{k} c(p, e, s)$$

This is the fraction of runs in which engine e mentions brand B in response to prompt p.

Per-engine MV (the atomic, cross-study comparable unit):

$$MV_e = 100 \times \frac{1}{n} \sum_{p=1}^{n} r_e(p)$$

MV_e is the mean mention rate for engine e across the full prompt set, expressed on a 0–100 scale.

Headline MV (equal weights over the declared engine list):

$$MV = \frac{1}{E} \sum_{e=1}^{E} MV_e$$

The headline reconciles exactly with its parts. It is the arithmetic mean of the per-engine scores. Equal engine weighting is used throughout this specification. The rationale is longitudinal comparability and engine neutrality. A score should change because AI mention behaviour changed, not because market shares among the declared engines shifted. Market-share-weighted overlays are permitted as documented non-canonical variants, and they must be clearly distinguished from canonical MV.

An MV of 54 means the brand appears in an estimated 54% of AI responses to category-relevant prompts, averaged across the declared engines.

4.3 Error quantification

Reported uncertainty on MV decomposes into four sources, each with a different operational treatment.

The first is **run stochasticity**. Within a single prompt-engine cell, $r_e(p)$ is a binomial proportion over k trials, and repeated runs at that cell quantify the engine's stochastic answer behaviour. The $k \geq 3$ floor is the minimum run count for any measurement, and the $n \times k \geq 150$ rule is the collection-size convention for published benchmark figures, informed by the binomial worst-case half-width (approximately ± 8 points per engine at $n = 30$, $k = 5$).

The second is **prompt-set sensitivity**. Because the prompt set is held fixed for the estimand (§4.4), the prompt-level bootstrap is best read as estimating how the headline MV would shift if the prompt set had been a comparable but different selection from the same generation methodology. Resampling prompts re-weights which prompts contribute, and the resulting interval reflects design uncertainty under a fixed estimand rather than sampling uncertainty in the conventional sense. The use of bootstrap to quantify topic-set sensitivity under a fixed evaluation design follows the information retrieval evaluation literature (Buckley and Voorhees, 2004; Sakai, 2014).

The third is **engine temporal drift**. Across scans, engine behaviour itself changes through model updates, retrieval-index updates, and ranking-policy updates. The framework treats drift as longitudinal signal rather than measurement noise, documents it through the version-marked series (§8.4), and isolates protocol artefact from drift through the control arm (§7.5).

The fourth, relevant for MA and MS, is **evaluator error**. Evaluator drift and inter-evaluator disagreement add uncertainty beyond MV's three sources. Human calibration (§7) and the evaluator-agreement reporting targets (§7.2) address this, and the LLM-as-judge bias literature (Zheng et al., 2023) names the position, verbosity, self-preference, and prompt-sensitivity risks the evaluator architecture must control.

n and k are tradeable against each other above the $k \geq 3$ floor, but they are not fully interchangeable. Increasing k reduces within-cell stochasticity, while increasing n improves coverage of the prompt space. Which to invest in depends on the dominant uncertainty source in the application context. The collection rule is a provisional planning convention pending calibration against observed score dispersion. The $k \geq 3$ floor applies to all measurement, including monitoring scans, but only scans feeding published benchmark figures must clear the $n \times k \geq 150$ threshold.

Engine differences are systematic rather than sampling noise. They reflect different retrieval architectures, training data, and prompt interpretation tendencies. Per-engine scores are reported always, and engine effects are analogous to house effects in survey methodology, where different instruments applied to the same phenomenon produce systematically different estimates (e.g., Jackman, 2005).

Rogan-Gladen correction. When mention detection has known false-positive and false-negative rates (for example, due to abbreviation ambiguity or entity resolution errors), the observed mention rate $r_e(p)$ can be corrected for imperfect detection using the method of Rogan and Gladen (1978):

$$r_{\text{corrected}} = \frac{r_{\text{observed}} + \text{specificity} - 1}{\text{sensitivity} + \text{specificity} - 1}$$

Application requires empirical estimates of detection sensitivity and specificity, typically derived from manual audit of a random sample. The uncorrected estimate is the default, and correction is documented when applied. Edge cases (sensitivity + specificity < 1, indicating a worse-than-random detector) are clamped to [0, 1] and flagged in the reporting conditions block; their occurrence signals a misconfigured detection rule rather than a parameter regime the formula handles.

4.4 Statistical properties

MV is defined over the declared, immutable prompt set. The estimand is the brand's mention probability across that fixed set, and as k increases MV converges to that fixed-set quantity. The framework does not claim MV as an estimate of mention probability over a hypothetical wider population of category prompts. Representativeness of the prompt set is a governance property (Appendix A), not a superpopulation claim.

The per-engine mean mention probability has an interpretable unit (expected mention probability for a randomly drawn prompt from the declared set) and aggregates linearly across engines into an interpretable headline score. The main alternatives each require additional analytical choices. Threshold intersection requires a fixed cut-point. Quantile statistics summarise distribution shape that reach and HHI already capture as companions. AUC presupposes a ranking over prompts that is not specified in this framework.

Asymptotic behaviour. As $k \to \infty$ over the declared prompt set of size n, the per-engine MV converges to the true per-engine mention probability over that set by the law of large numbers. When the Rogan-Gladen correction is applied, the corrected estimate is asymptotically unbiased given correctly specified sensitivity and specificity. The binomial worst-case half-width used as the collection-planning heuristic (§4.3) holds well at $n \times k \geq 150$, and the reported interval is computed by prompt-level bootstrap and accounts for the clustered structure that the binomial figure does not.

Finite-sample sensitivity. The $n \times k \geq 150$ threshold addresses the finite-sample regime. Below this threshold, the binomial variance term dominates and estimated CIs are unreliable as movement guides. The reach companion statistic is more sensitive to sampling than the headline MV. Reach estimated from $n < 20$ prompts should be interpreted with additional caution, as single unmentioned prompts have a disproportionate effect on the reach percentage.

4.5 Companion statistics

Three statistics carry the breadth and concentration information that MV, as a level measure, does not. They are reported alongside the headline, never merged into it.

- **Reach:** the percentage of prompts on which the brand is mentioned at least once, per engine and overall. A brand with MV 40 achieved by moderate presence across all prompts differs from one with the same MV achieved by high presence on a narrow slice.
- **Intensity:** for each engine, the mean mention rate among the prompts that engine reaches. The headline intensity is the equal-weighted mean of per-engine intensities, following the same per-engine-first aggregation pattern as MV (§4.2). Together with reach, intensity describes the shape of the mention rate distribution.
- **Concentration:** the Herfindahl-Hirschman Index (HHI) computed over prompt-level contribution shares (Rhoades, 1993). High concentration with a given MV signals fragile, narrow visibility concentrated on a small number of prompts. HHI is selected over the Gini coefficient and coefficient of variation because it has an interpretable uniform baseline (1/n), so a multiple of uniform is the immediately readable quantity (in the worked example, approximately 1.4× uniform concentration).

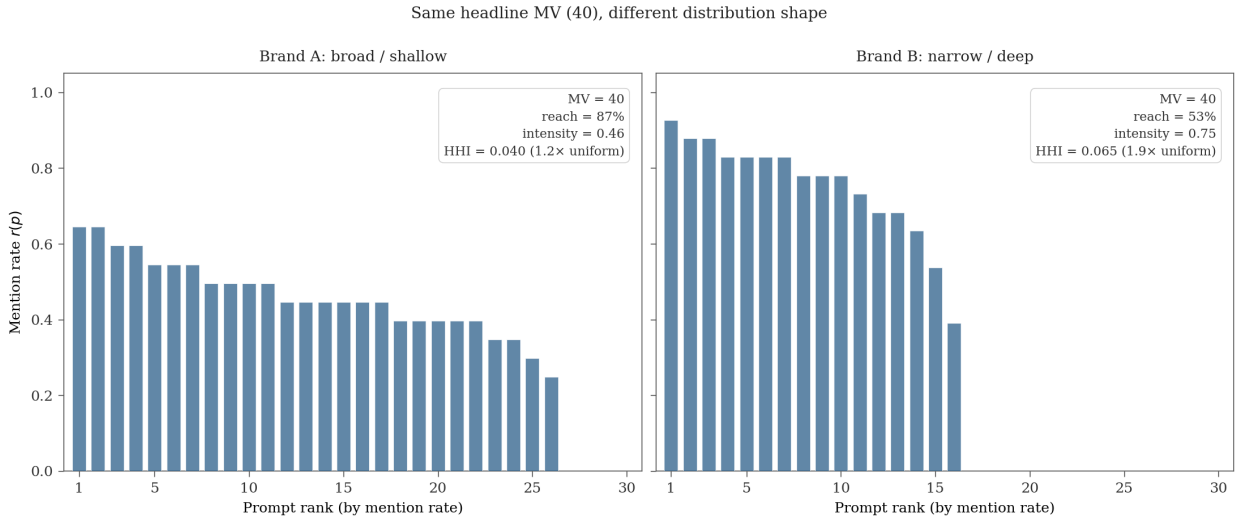


Figure 2: Two brands with identical headline MV (40) but different mention distributions.

Left: broad reach, lower intensity (many prompts at moderate mention rates, HHI close to uniform). Right: narrow reach, higher intensity (few prompts at high mention rates, HHI elevated). Companion statistics

distinguish these states, while headline MV alone does not. Data: synthetic illustration.

4.6 Worked example

For brand Acme Analytics, n = 30 prompts, E = 5 engines (ChatGPT, Claude, Perplexity, Gemini, Microsoft Copilot), k = 5 runs, with n × k = 150 per engine, meeting the benchmark publication threshold.

Engine	Per-engine MV
GPT	62
CLD	58
PPX	46
GEM	54
COP	40

Headline MV = (62 + 58 + 46 + 54 + 40) / 5 = 260 / 5 = **52**.

Companion statistics on the same sample are reach 87% (per-engine 67–87%), headline intensity 0.68 (per-engine 0.57–0.73), and HHI 0.047 against a uniform baseline of 0.033. The full per-prompt per-engine derivation is in Appendix B; full reporting conditions for a published score are specified in §8.2.

---

5. MA: brand-mention alignment

> **MA is not an external fact-checking score.** MA measures the consistency between AI brand descriptions and a brand-provided, version-controlled Brand Profile. It is a representational-consistency measure, not a measure of factual truth against an external corpus. The construct boundary is treated formally in §5.2 and §10.2.

5.1 Definition

MA is the expected alignment given mention, namely the mean alignment score across all eligible mentioning cells over the full prompt set.

An MA of 68 means the average eligible mentioning cell receives an alignment score of 68/100 against the Brand Profile.

MA is expressed 0–100. The denominator is eligible mentioning cells, not total prompts. Coverage (eligible cells as a proportion of total mentioning cells) is reported alongside every MA score.

5.2 Ground truth: the Brand Profile

MA requires a version-controlled ground-truth document, the **Brand Profile**, specifying the facts against which AI descriptions are evaluated. The Brand Profile is provided by the brand and held constant across measurement periods. Version control applies. Any update requires a new version identifier and a marker on the MA series flagging that the reference standard has moved (see §8.4). 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.

MA measures representational consistency, not epistemological truth. The Brand Profile is the operational ground truth, and MA measures conformance to it. A brand with an incomplete or strategically framed Brand Profile receives MA scores reflecting alignment with that declared profile.

5.2.1 The three-tier claim model

Brand Profile fields are classified into three tiers:

Tier	Nature	Examples	MA treatment
Externally verifiable facts	Objectively checkable against public sources	Founding year, company type, product category, pricing	Full weight
Brand-stated facts	Asserted by the brand; operationally authoritative	Feature descriptions, use cases, named integrations	Full weight
Positioning claims	Interpretive or contestable	"Best for enterprise", "leading platform"	Excluded from scoring

MA scores reflect alignment on Tiers 1 and 2 only.

Brand Profile Tier 1 minimum (normative). A Brand Profile used for benchmark MA publication must contain at least three Tier 1 fields (externally verifiable facts). The requirement sits in the same family of structural constraints as prompt-set immutability (§8.5). Tier 1 fields are the only externally auditable component of MA's ground truth, since a Brand Profile without externally verifiable fields is unauditable and the MA score it produces is unanchored to any external reality. The Tier 1 minimum is the framework's convergent-validity floor, and the full convergent-validity argument is in §10.2.

Field types and scoring grades are published in full in Appendix C.

5.3 Formal definition

Let B be the brand, F_B the Brand Profile, P the full prompt set of n prompts, E the declared engine set, and k the run count per prompt-engine pair.

For each prompt p and each engine response that mentions brand B, let $A(r, F_B) \in [0, 1]$ be the equal-weighted alignment score of response r against F_B, computed by the evaluator using the published rubric in Appendix C. ("Weighted" is reserved for the documented non-default variant in which fields receive differential weights.)

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). Responses that mention the brand name incidentally, without descriptive claims, are excluded. This eligibility rule ensures near-zero-information observations do not distort the score.

Cell-level eligibility (mixed runs). Within a prompt-engine cell, identify the mentioning runs and the subset that meet the MA eligibility criterion above. If none qualify, the cell counts as a mentioning but ineligible cell, contributing to coverage but not to MA. If one or more qualify, the cell-level MA value is the average alignment score across the eligible mentioning runs, and the cell contributes one unit to the per-engine MA denominator. The same rule applies to MS in §6.3.

Per-engine MA: the mean alignment score over that engine's eligible mentioning cells across all prompts.

Headline MA: the mean of per-engine MA values over engines meeting the eligible-cell floor below.

Eligible-cell floor. A per-engine MA based on fewer than 10 eligible mentioning cells is excluded from benchmark tables, rankings, and headline aggregation. It may appear in full diagnostic reporting, flagged as indicative, with eligible count and confidence interval attached. The rationale is straightforward. At $N = 10$ and realistic score dispersion (SD ≈ 0.25), the 95% CI half-width is approximately ± 15 points, which is wide but interpretable. Below 10, the interval spans most of the scale. The value 10 sits at the conservative end of small-cell suppression practice in official statistics. It is a working assumption pending calibration.

The mean alignment score is preferred over a binary threshold (aligned/not aligned) because it preserves the graded information in the rubric and supports interval-level comparisons across Brand Profile versions.

The trigger is the eligible count, not MV. A healthy MV does not guarantee an adequate MA denominator. Many mentions may be bare-name list inclusions with no descriptive content.

Coverage reporting. Low MA coverage is itself a finding. If most mentions are bare-name 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. It differs from both low visibility and low alignment. Coverage below 50% is treated as a caution condition. The score is reported, but its interpretation should note that a majority of mentions are bare-name or list mentions without substantive description.

5.4 Interpretation and worked example

For Acme Analytics, continuing the example from §4.6:

Across 114 mentioning cells in total, 70 were eligible (triggering at least two Brand Profile fields; coincidentally equal to the MS eligible cell count in this illustration, see Appendix B Part 3). MA coverage = 70 / 114 = **61%**.

Alignment scores by engine across eligible mentioning cells:

Engine	Eligible cells	Per-engine MA
GPT	16	71
CLD	15	76
PPX	12	63
GEM	16	68
COP	11	59

All five engines clear the 10-cell floor, so all five contribute to the headline.

Headline MA = (71 + 76 + 63 + 68 + 59) / 5 = 337 / 5 = **67**.

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. MS: brand-mention sentiment

6.1 Construct

MS measures sentiment toward the brand in AI-generated responses, namely whether the response speaks positively or negatively about the brand, not the overall tone of the text it appears in.

A response that is broadly critical of a category but singles out the brand approvingly is positive for the brand. A response that is generally positive 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, and implementations diverging on this definition produce silently non-comparable scores. The full definition with worked counter-examples appears in Appendix D.

6.2 Coding method

Each eligible brand-directed characterisation within a mentioning cell is classified as positive, neutral, or negative by the evaluator against the published codebook in Appendix D before cell-level aggregation. The eligibility machinery mirrors MA. A minimum-content rule applies for sentiment evaluability. The evaluator

is an LLM from a different model family than the evaluated engine, calibrated against a human gold standard (see §7).

Inter-rater reliability is reported with Krippendorff's alpha, consistent with the MA validation design. One reference to content-analysis methodology (Krippendorff, 2004) covers the coding standard.

MS is an interpretive measure. Sentiment judgements are less determinate than factual alignment checks. The expected reliability ceiling for MS is below MA's, and the paper states differentiated expected agreement levels for the two measures rather than implying equal precision.

Shares-plus-net is preferred over averaged bipolar scores because it distinguishes polarised from uniformly neutral distributions (Figure 3), a distinction that a single net score collapses. The polarised-vs-uniformly-neutral failure mode is the structural reason the framework reports shares alongside the net rather than collapsing to a single bipolar score.

6.3 Reporting: shares plus net

The three shares are reported per engine and overall: % positive, % neutral, % negative.

Cell-level eligibility (mixed runs). The §5.3 rule applies to MS with the eligibility criterion swapped. An MS-eligible run contains at least one brand-directed evaluative characterisation, and the cell-level MS classification is constructed from the MS-eligible mentioning runs in the cell. The MS eligible-cell set may therefore differ from the MA eligible-cell set on the same scan.

The eligible-cell floor from §5.3 applies to per-engine MS identically. At the floor of 10 eligible mentioning cells, shares move in 10-point steps, which the mandatory confidence intervals make visible.

Per-engine shares and per-engine net are computed first. The headline shares and headline net are the equal-weighted mean of per-engine values over engines clearing the eligible-cell floor (the same aggregation architecture as MV and MA).

The headline figure is **net sentiment = % positive − % negative**, on a -100 to +100 scale. The signed scale is the appropriate one for this construct, and the asymmetry with MV and MA scales (which are both non-negative) is noted here and does not represent an inconsistency.

The shares-based construction distinguishes two conditions that averaged bipolar scores collapse: a high neutral share (most descriptions neither favourable nor unfavourable) and high positive plus high negative shares netting near zero (the brand praised and criticised in roughly equal measure). Both produce a net near zero but represent different remediation problems.

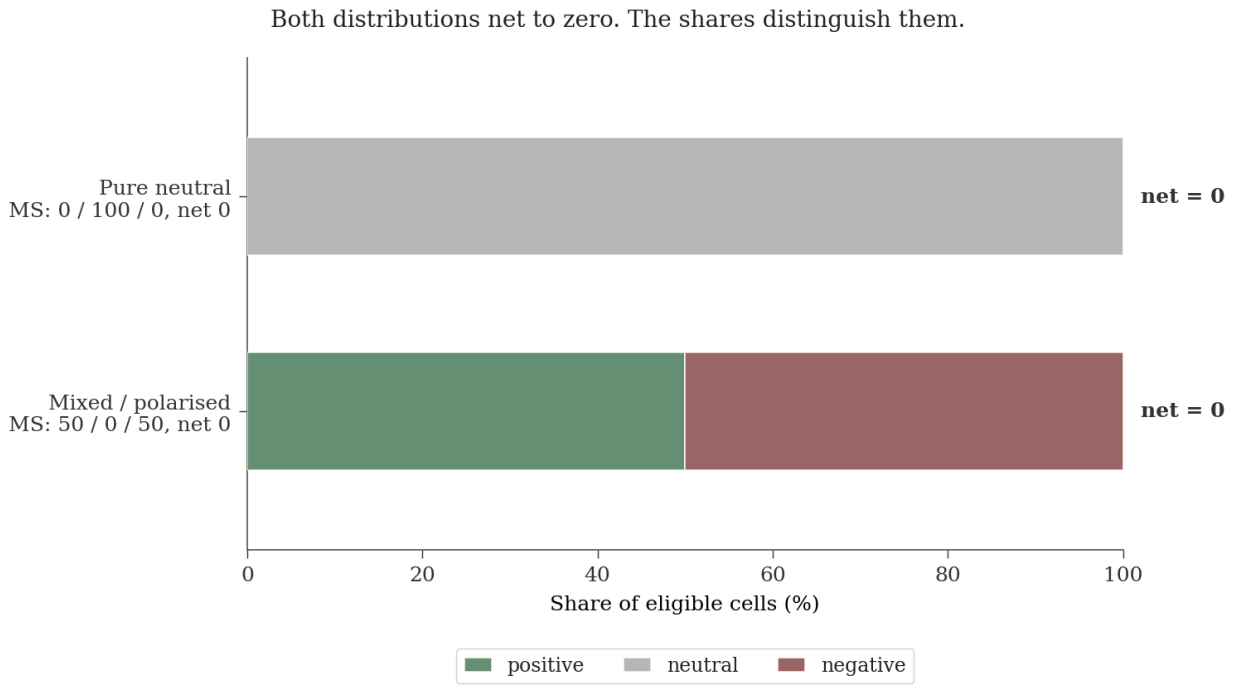


Figure 3: Two sentiment distributions that both net to zero.

Top: 0% positive, 100% neutral, 0% negative (pure neutral). Bottom: 50% positive, 0% neutral, 50% negative (mixed/polarised). Averaged bipolar scores collapse both; shares-based reporting distinguishes them. Data: synthetic illustration.

Display bands for the featured figure are derived from shares and net, evaluated in precedence order: Mixed (positive $\geq 25\%$ and negative $\geq 25\%$); Positive (net $\geq +20$); Negative (net ≤ -20); Neutral otherwise. The bands are presentation conventions for colour encoding, declared in figure notes; the underlying shares and net remain continuous. The ± 20 threshold exceeds the approximate ± 15 CI at the eligible-cell floor, so a colour assignment is unlikely to be a pure sampling artefact.

At the eligible-cell floor of 10, the Mixed band classification (positive share $\geq 25\%$ and negative share $\geq 25\%$) may be unreliable. At 10 eligible mentioning cells, a 30% share is 3 cells, so the band trigger is sensitive to single-cell movements. Mixed band classifications drawn at the floor require either elevated k or human review before downstream use.

A net sentiment of +24 (52% positive / 20% neutral / 28% negative) means that, when AI describes the brand, positive characterisations outnumber negative ones by 24 points of share.

6.4 Known risks

LLM evaluators exhibit a documented positivity lean in annotation tasks. The human calibration step must test for a systematic offset on brand-directed sentiment specifically, and correct it before production use. A positivity-offset test is included in the validation protocol (§9).

Aspect-level sentiment disaggregated by brand attribute (pricing sentiment, feature sentiment, support sentiment) is a diagnostic extension parallel to field-level MA disaggregation. It is not developed in the present specification.

6.5 Worked example

For Acme Analytics, 70 eligible mentioning cells across five engines:

Engine	Eligible cells	Positive	Neutral	Negative	Net
GPT	16	50%	25%	25%	+25
CLD	15	60%	20%	20%	+40
PPX	12	42%	25%	33%	+9
GEM	16	50%	25%	25%	+25
COP	11	55%	18%	27%	+28

Headline shares (equal-weighted mean of per-engine shares): 51% positive / 23% neutral / 26% negative. Headline net (mean of per-engine nets) = (25 + 40 + 9 + 25 + 28) / 5 = **+25**.

Display band: Mixed (positive 51% $\geq$ 25% and negative 26% $\geq$ 25%).

The positive and negative shares carry substantial weight. The brand receives balanced characterisations, praised for integration capability and ease of use, criticised for pricing ambiguity in AI responses. The Mixed band at net +25 illustrates the precedence rule. The mixed condition triggers before the positive condition even when net sentiment is above zero, because both favourable and unfavourable characterisations are well-represented. Sentiment-driver analysis (representative excerpts by codebook category) appears in the full diagnostic report.

---

7. Evaluator Architecture

Both MA and MS are LLM-scored in production and human-calibrated. The evaluator is a measurement instrument, calibrated, not trusted.

7.1 Gold standard

Once per instrument version, a stratified sample of responses (200 responses, by default) is double-coded by human annotators against the published rubric (for MA) and the published codebook (for MS). Disagreements are adjudicated to produce a single adjudicated label per response. Human-human agreement before adjudication is the reliability ceiling, and the LLM evaluator's target is agreement with adjudicated humans comparable to that ceiling.

7.2 Instrument validation

Before any score is published from an evaluator, the following targets must be cleared:

- Cohen's kappa (for categorical field verdicts) or Krippendorff's alpha (for graded fields): target ≥ 0.75.
- LLM-human score correlation: r ≥ 0.80.
- For MS specifically, a positivity-offset test, checking whether the LLM evaluator systematically over-classifies responses as positive relative to the human gold standard.

Scores are not published from an evaluator that has not cleared validation.

The validation targets ($\alpha \geq 0.75$) apply to both instruments. The expected human–human ceilings differ. Pre-adjudication human agreement on MS is expected to fall measurably below MA's, reflecting the greater interpretive latitude in sentiment attribution relative to fact-checking. Both ceilings are reported alongside the LLM evaluator scores, and the evaluator is assessed against its own instrument's human ceiling, not a fixed absolute.

7.3 Production

The production evaluator is an LLM from a different model family than the engine being evaluated. An evaluator from the same model family may reproduce the same training-data errors as the evaluated engine,

artificially inflating agreement scores through shared pre-training bias rather than genuine correctness. Cross-family selection reduces this risk. Residual risk from overlapping pre-training corpora across different model families exists and is handled by the human calibration protocol.

Where the standard engine set includes a model from the evaluator's own family, responses from that engine are scored by a designated second evaluator from a different family, validated against the same gold standard. If no second evaluator is operationally available, same-family scoring is permitted only with (i) explicit flagging in the conditions block, and (ii) an elevated human-audit sample for that engine's responses. This condition is structural, not an edge case. Any standard engine list spanning major AI providers will include at least one same-family engine.

The production evaluator runs as an ensemble: median of three independent evaluations per response. The evaluator model version is pinned and documented in all published reporting. An unpinned evaluator is an instrument that silently recalibrates, and this specification prohibits it.

The evaluator session is independent from the data collection session. The evaluator must not have prior exposure to the responses being evaluated within the same session.

7.4 Drift control

A small human re-validation sample is evaluated each measurement cycle to detect evaluator drift. On any evaluator model change, full re-validation against the human gold standard is required before new scores are published. The evaluator version documented in the reporting conditions makes drift detectable across historical records.

7.5 Control-arm architecture and recalibration protocol

The framework requires a version-pinned control arm designed to minimise engine-side drift: a locally hosted open-weight model whose response behaviour is stable by construction (nominally zero-drift, with residual variation from hardware and inference-framework factors acknowledged in §10.1). The control arm's role is to validate the noise model, not to measure a noise level.

Reference model. The reference implementation specifies Llama-3.1-8B-Instruct at a pinned commit hash (recorded in the conditions block alongside the evaluator version), operated with temperature 0, a fixed pseudo-random seed, and a stated quantisation level (FP16 or INT8, whichever the implementation pins). Implementers may substitute a different open-weight model provided the substitution is logged and the candidate passes the determinism test below. Any change to the reference model is an instrument change and triggers full recalibration.

Determinism test. Floating-point arithmetic varies across hardware platforms and inference frameworks. The control arm is not required to produce bit-identical output across deployments. It is required to produce statistically stable output. With the same model, same seed, same prompt, and same temperature, response sequences must show token-level disagreement below a stated threshold. Defaults: $\leq 1\%$ of tokens differ across runs on a single hardware platform; $\leq 5\%$ of tokens differ across two hardware platforms at the same quantisation level. A deployment failing either threshold is not admitted as a control arm. The test must be passed and logged before any control-arm score is published, and re-passed after any infrastructure change. The thresholds are working assumptions pending calibration against observed cross-platform variance.

Purpose. Per-scan confidence intervals are computed from that scan's own observed rates, and engine-side stochasticity is measured live on each commercial engine through within-scan k-run replication. Both self-update every scan. What cannot be checked from commercial scans alone is whether the measurement protocol itself (harness, run spacing, parsing, mention-detection rules) injects variance that the binomial model does not account for. On commercial engines, protocol artefact and engine behaviour are confounded. On a pinned control, the engine contributes nothing to the variance, so any excess over the model's prediction isolates protocol artefact. This validation licences the per-scan noise arithmetic that calibrates alert thresholds on commercial surfaces, the noise floor behind the "real movement versus measurement wobble" judgement that every practitioner alert depends on.

Fencing. The control arm's MV, MA, and MS values describe no buyer-facing surface and are excluded from all brand-level inferential tests, including the incremental-validity correlations in §9.3, any published brand

scores, and any competitive maps. Its data serves three purposes only: noise-model validation, additional evaluator-calibration material, and replicability demonstration (the control arm is the one engine any third party can rerun exactly).

Canary. Between full recalibrations, a low-k control scan runs on a stated cadence (quarterly by default) as a pipeline canary. A pinned model's score is statistically flat by construction, so any measured movement on the canary is evidence the measurement pipeline changed, not the world. Escalation rule: a canary anomaly triggers immediate full-grade recalibration and a hold on published alerts until the source of movement is identified and resolved. The canary is deliberately underpowered (it detects gross pipeline breaks, not subtle model misfit) and is explicitly not a recalibration.

Recalibration triggers. Full-grade recalibration (benchmark-grade n × k, full statistical analysis) is required on (i) any material change to the measurement instrument, including n or k tier changes, harness modifications, parsing rule updates, mention-detection changes, or infrastructure migration; (ii) canary escalation; (iii) the backstop interval described below.

Backstop interval. A scheduled full-grade recalibration exists solely to catch silent instrument decay too subtle for the canary, a mode the canary is not designed to detect. The backstop interval cannot be derived analytically, because the staleness rate of silent infrastructure change is not known in advance. It is therefore entered in the framework's working-assumption register as an operational convention. The initial value is 12 months, self-revising on its own evidence. Sustained clean backstop results and clean canary history support lengthening the interval. A backstop that catches a misfit the canary missed supports shortening it, and upgrading canary sensitivity.

Rolling clock. The backstop clock resets on the most recent full-grade control run, regardless of what triggered it. Event-triggered recalibrations reset the clock, while canary runs never do. A recalibration triggered by an instrument change in month 11 makes the next backstop due in month 23, not month 12. Without this rule, routine canary runs could be argued to reset the clock, deferring the rigorous full-grade run indefinitely.

---

8. Comparability and Governance

8.1 Declared conditions

Every externally published score carries the following conditions block. Published means externally reported figures in benchmark reports, papers, or marketing claims. In-product displays must make the same conditions accessible.

Mandatory reporting conditions block:

- Engine list (using the three-character codes defined in §8.3)
- Prompt set version identifier
- Measurement period (start and end dates or year-month)
- n (number of prompts), k (runs per prompt-engine pair)
- Evaluator model and version (for MA and MS)
- Eligible cell counts and MA/MS coverage statistics

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

A score without its conditions cannot be compared to another score. Conditions are part of the result.

8.2 Per-engine comparison rules

The atomic comparable unit is the per-engine score. Two studies measuring the same engine over the same period compare directly on that engine's score. Studies with different engine sets compare on the overlap. Headline scores are comparable only when the declared engine lists match.

Engine model changes beneath a declared surface (e.g., a ChatGPT model version update that changes mention behaviour) flow 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, and is not an artefact to adjust away.

8.3 Standard engine list

The paper specifies the selection rules. The current membership lives in Aiviara's published reference materials, where it tracks the market without requiring a specification change.

Selection criteria (four, all required): (1) a buyer-facing answer surface; (2) measurable by repeatable programmatic querying; (3) usage above a declared threshold among the buyer population the list serves (for a B2B-oriented list, enterprise-seat prevalence counts alongside consumer traffic), per referenced third-party usage data; (4) the resulting set spans both mention mechanisms, training-data and retrieval.

Scope statement (applies to all engines uniformly): the framework measures each engine's public, web-grounded surface. For category-relevant prompts, this proxies the discovery-relevant behaviour of licensed enterprise variants. Internal-data-grounded responses are out of scope for all engines equally.

Current membership and code convention. Six engines currently meet the criteria: ChatGPT, Gemini, Google AI Overviews, Perplexity, Claude, and Microsoft Copilot. AI Overviews and Gemini are separate engines with the same model family but materially different surfaces and retrieval behaviour. Engine codes are three characters derived from the marketed surface name (GPT, CLD, COP, GEM, AIO, PPX), extended on collision. A list identifier is the codes sorted alphabetically and hyphen-joined; the current standard is **AIO-CLD-COP-GEM-GPT-PPX**.

A different engine set automatically produces a different identifier. Mismatched lists cannot silently masquerade as comparable. The full engine list always appears in the mandatory conditions block, and codes are display compression.

8.4 Temporal series integrity

Scores are single-scan-window quantities. The temporal view is the dated series of unmodified scores. No score is ever adjusted. Changes are marked, decomposed, and explained.

Engine surface changes (world changes): model updates behind a declared surface flow through the series as real movement. 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 a prompt set is updated (category-change trigger only; see Appendix A), both versions run 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. The decomposition states how much of the difference is movement on familiar questions (prompts in both versions) and how much is position in the new question space (prompts in the new version only). An example is "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 then continues with new-set scores. The old-set history stands as measured, and no splicing or rescaling occurs.

Evaluator changes: handled by mandatory re-validation against the human gold standard. No score adjustment.

Brand Profile version changes: a marker on the MA series only. MV does not condition on the Brand Profile (the brand was mentioned, or it was not), so the MV series runs unbroken and unannotated through the event. The MA series likewise continues through the marker as a single series, and what it carries at the boundary is a caution, not a severance. MA either side is measured against different declared facts (the reference standard for "correct" has moved), so the cross-boundary MA gap is not directly interpretable as performance change. The practitioner reads movement across the marker knowing the reference moved.

This marker is stronger than the prompt-set marker, and stating why is the point. 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 MA reference standard itself moves, so the cross-boundary gap is not directly interpretable as performance. The framework distinguishes

the two transitions because they call for different practitioner readings, not because one is a "hard" event and the other "soft".

No single adjustment factor purporting to make arbitrary studies comparable is permitted anywhere in the specification.

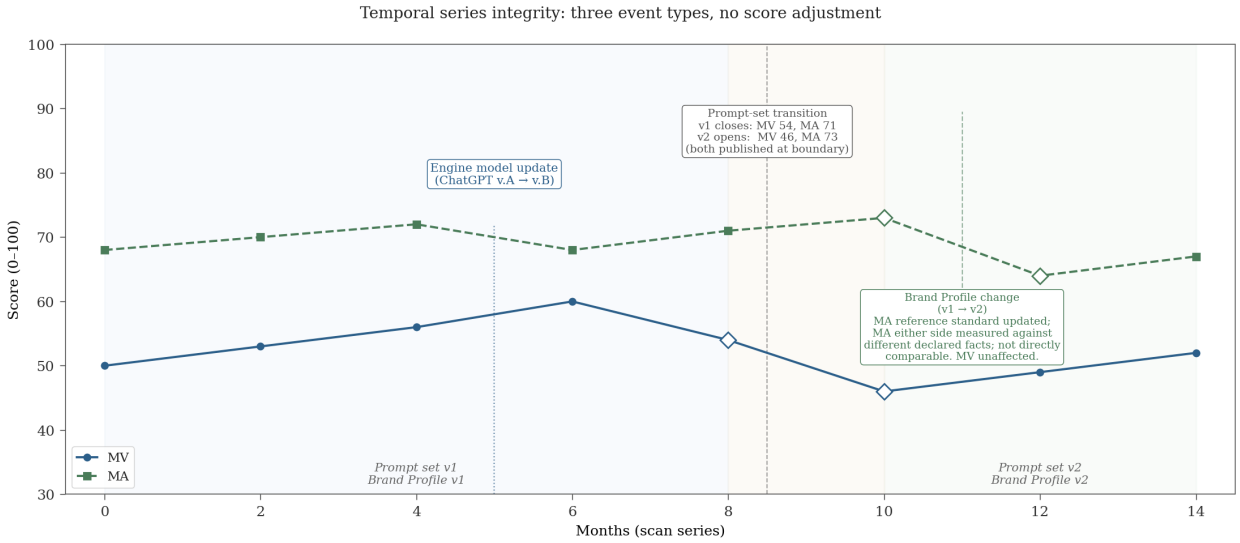


Figure 4: Temporal series integrity.

Three event types under the no-adjustment principle: engine model update (real movement, no break), prompt-set transition (both scores published at the boundary, decomposable into movement on familiar questions vs position in the new question space), and Brand Profile change (MA reference standard moves, so MA segments either side are not directly comparable; MV is unaffected because visibility does not condition on the Brand Profile). No score is adjusted, and the MV series runs continuously throughout. Data: synthetic illustration.

8.5 Prompt set governance

Prompt set immutability follows the test-collection principle in information retrieval, under which a fixed, reusable stimulus set enables reproducible comparison across studies and over time (Voorhees & Harman, 2005). Prompt sets are immutable within a version. A prompt set may only be updated when there is a documented change in the category's scope or a category becomes obsolete. Each update requires a logged break in the time series with the version marker and overlap-window protocol described above.

Practitioners may track additional prompts for diagnostic purposes beyond the standardised benchmark set. Custom sets must be maintained separately, never merged with the standardised set, and clearly labelled as non-comparable to MV benchmark figures.

Practitioners implementing this framework must review the terms of service of each commercial engine before conducting systematic automated querying at the collection scales described.

8.6 Manipulation resistance

The framework's manipulation resistance derives from governance, not from formula structure alone.

MV. A brand cannot inflate its score by adding prompts on which it performs well, because the canonical prompt set is held fixed by the measurement operator. Narrow optimisation resistance is a property of the mean. Improving mention frequency on any prompt raises MV. Brands have no incentive to concentrate effort on a small subset of prompts, because every prompt contributes equally.

MA. Prompt set governance inherited from MV means MA cannot be inflated by selecting prompts on which the brand is well-described. Brand Profile governance applies in addition. A party that weakened alignment

criteria by modifying the Brand Profile could inflate MA scores. The Brand Profile must be version-controlled with the same discipline as the prompt set.

An adversarial vector exists. A brand could craft a Brand Profile that agrees with AI's known misrepresentations, achieving high MA without improving actual accuracy. Brand Profile version control and publication make the profile auditable across measurement periods, and the residual risk (that a strategically framed profile systematically agrees with AI errors) is acknowledged plainly here.

A second vector arises when content is engineered so that AI mentions trigger only easy-to-satisfy Brand Profile fields. This vector gains no advantage, because the two-field eligibility threshold means a mentioning cell must trigger at least two fields to enter scoring, and field-level disaggregation makes selective-triggering patterns visible in diagnostic reporting.

Content-ecosystem manipulation is the framework's hardest adversarial case. A brand that floods the web with synthetic third-party content will, if that content is indexed and trained on, achieve genuine increases in mention frequency. MV will correctly reflect the changed information state, even if the change was manufactured. Governance controls cannot prevent this. They can only make the measurement accurate. Sustained score inflation without corresponding improvement in independently verifiable brand representation is the monitoring signal for adversarial pressure.

Three further vectors are harder to govern against than content-ecosystem manipulation.

Prompt poisoning. Seeding online content to cause AI engines to generate brand-mentioning prompts shifts what queries the measurement captures rather than how the brand scores on a fixed set. The prompt governance protocol (Appendix A) mitigates this through agent-generated prompt sampling from stated category criteria, but cannot eliminate it. Sufficiently large-scale content seeding may shift the category's query distribution itself.

Retrieval manipulation. Exploiting search-engine rank to force brand mentions into retrieval-augmented responses independently of training data affects the detection stage (MV) directly and may be undetectable without comparing retrieval-augmented and training-data-only engine responses.

Structured-data gaming. Injecting schema markup to make AI engines parse brand descriptions in ways that selectively satisfy Brand Profile fields affects MA specifically. Field-level score disaggregation and independent Brand Profile publication provide partial governance, but structured-data injection at scale is not auditable under the current framework.

---

9. The Generative Brand Mention Framework, the Featured Diagnostic, and the Practitioner Layer

9.1 Framework structure

The three metrics form a two-level structure:

Level	Metric	Question	Conditioning
1	MV	Does the brand appear?	Unconditional over the prompt set
2	MA	When it appears, is the description correct?	Given mention
2	MS	When it appears, is the characterisation positive or negative?	Given mention

MV, MA and MS are separable measures. Each addresses a distinct property (presence, descriptive accuracy given mention, and characterisation given mention), 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. Their empirical relationship is a validation question, addressed by the incremental-validity tests below.

9.2 The featured diagnostic: the Visibility–Alignment map

The featured diagnostic is the Visibility–Alignment map, with MS encoded as point colour using the MS display bands (§6.3). MA is placed on the horizontal axis and MV on the vertical, because MV is the more consequential dimension (visible brands can carry the larger practical discovery signal), and the diagnostic gradient therefore reads top-to-bottom for visibility and left-to-right for alignment. A single scan scores every brand in the category from the same collected responses, so the map is a competitive map by default. An MV of 54 is informative only against where 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

Quadrant boundaries at MV = 50 and MA = 50 are presentation conventions, declared as such. They are not metric thresholds. Because separability of MA from MV is a measured rather than assumed property (§9.1), the off-diagonal quadrants are diagnostic to the extent the validation study finds MA and MV non-redundant. The redundancy tests are in §9.3.

The diagnostic value of the map is read from position and colour together. A brand's full state is its quadrant combined with its sentiment band. The named states below are that combination.

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 brand stands when an answer engine assembles candidates: AI Champion (positive, the engine advocates for the brand), AI Contender (neutral, present in the consideration set without advocacy), AI Wildcard (mixed, on the list but dividing opinion, praised and criticised in roughly equal measure), AI Pariah (negative, characterised unfavourably). A position prefix names how the brand's standing departs from the visible-and-aligned ideal: none when the brand is both visible and accurately described; Undiscovered when accurate but unseen; Misrepresented when visible but inaccurately described; Misrepresented and 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, and 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. Each root is a shortlist-standing term, naming the state a brand holds when an engine compiles candidates, so a full label reads directly. A Misrepresented AI Wildcard is visible, inaccurately described, and divides opinion. Reading across a row holds position fixed and varies reception, while reading down a column holds reception fixed and varies position.

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. The §6.3 band definitions and precedence are unchanged. Mixed where positive $\geq 25\%$ and negative $\geq 25\%$; otherwise Positive where net $\geq +20$; Negative where net ≤ -20; Neutral otherwise. There is no merging of neutral with positive. A neutral mention yields AI Contender, a distinct state from the positive AI Champion.

From scores to descriptor. A brand's three scores resolve to one brand-representation state by two lookups. Position fixes the prefix (MV and MA each read against 50 place the brand in a quadrant, which sets the prefix), and sentiment fixes the root (the MS display band sets the reception root). The descriptor is prefix plus root. The teaching grid (Figure 5) shows the full taxonomy of sixteen states, and the competitive map (Figure 6) plots brands by score with the focus brand's descriptor annotated.

The sixteen-state lexicon is a presentation convention derived from the three measures, not a separately validated construct. The proposed validation programme (§9.6) is designed to test MV, MA, and MS. The states are a labelling layer on the same scores, useful for communication and per-brand reporting.

Insufficient data and AI Absent. A brand that clears the eligible-cell floor (§5.3) 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.

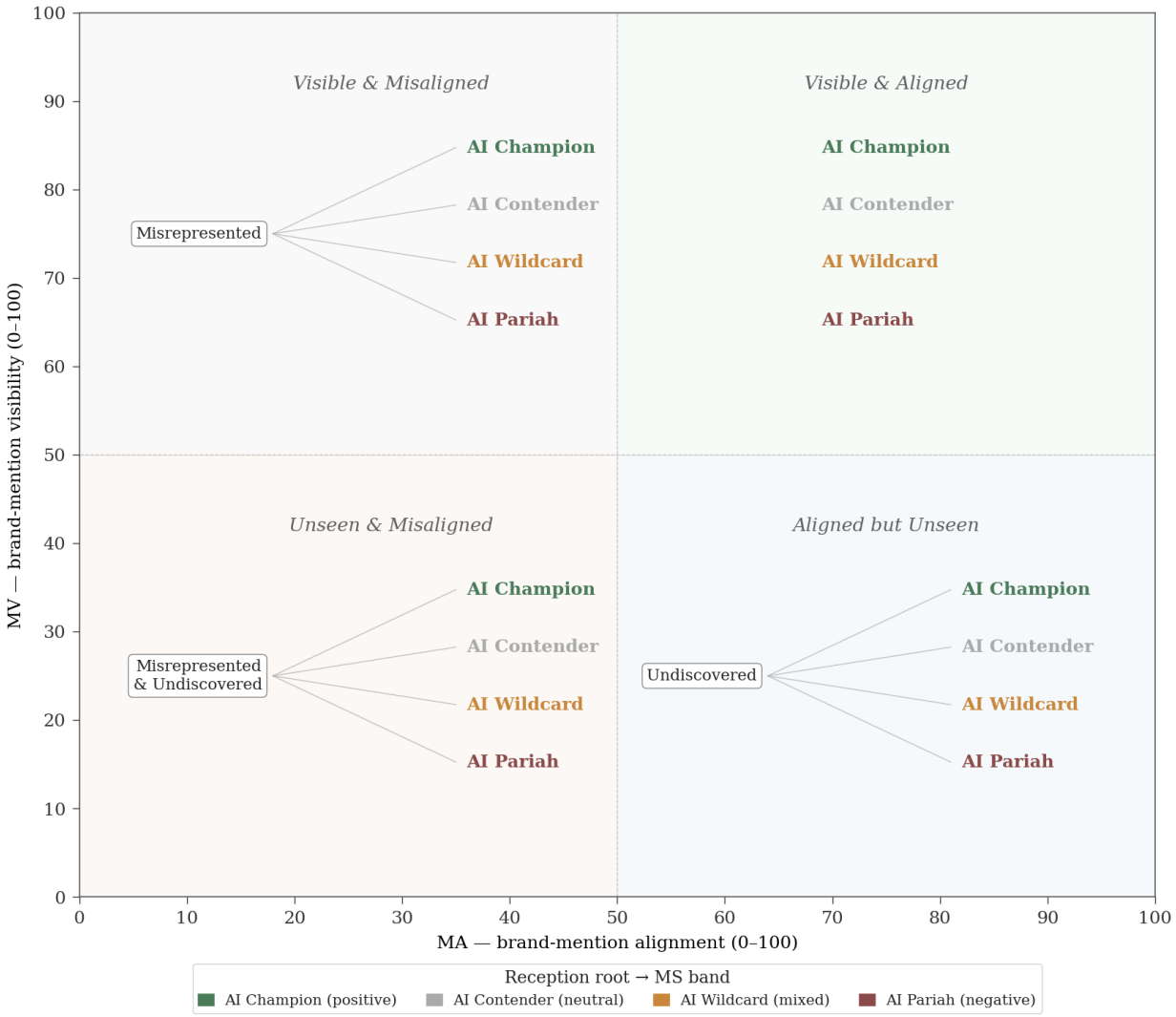
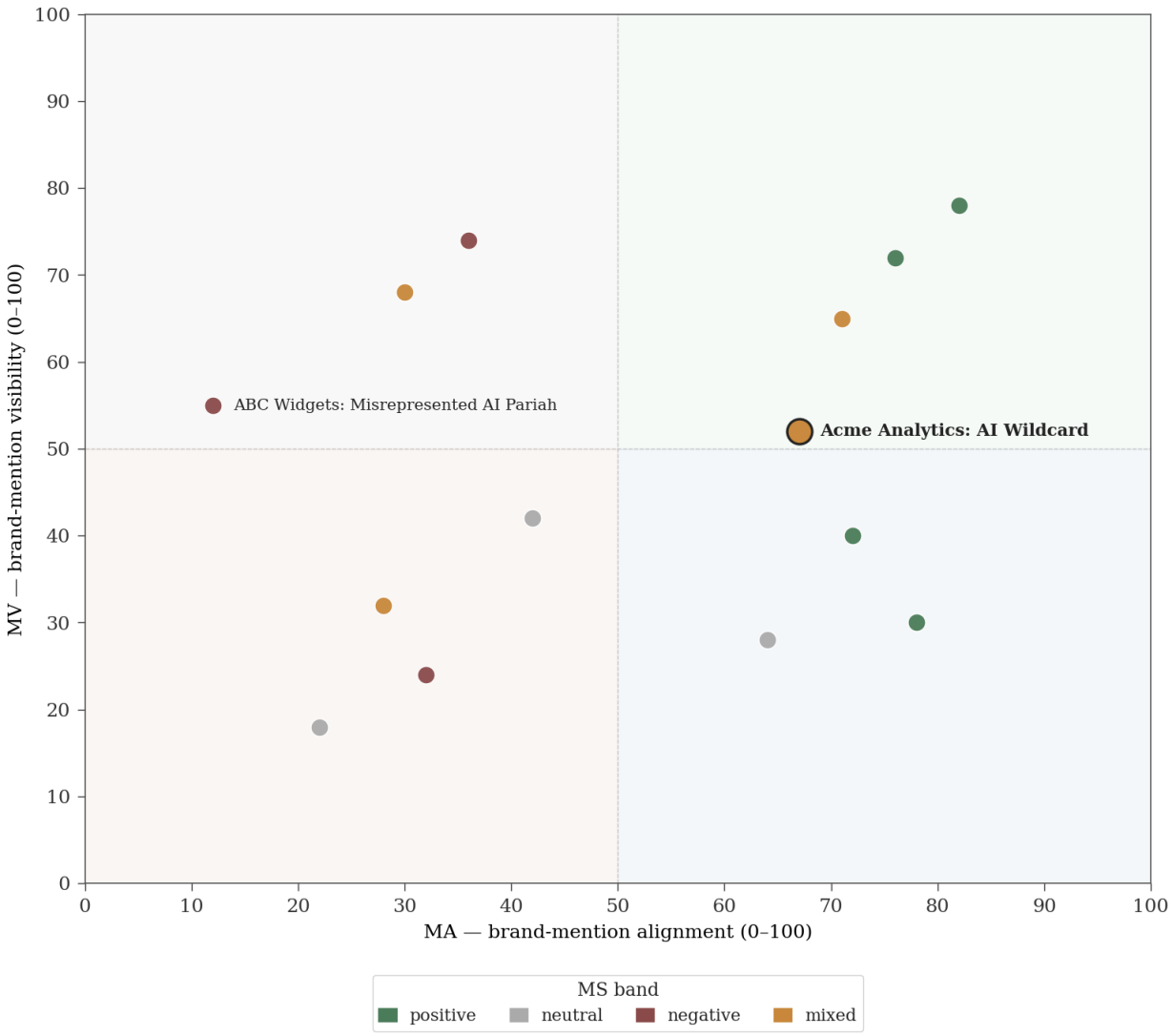


Figure 5: Teaching grid for the brand-representation taxonomy.

Each quadrant of the Visibility–Alignment plane carries a prefix node (none / Misrepresented / Undiscovered / Misrepresented and Undiscovered) and four reception-root branches, one per MS display band (AI Champion positive, AI Contender neutral, AI Wildcard mixed, AI Pariah negative). A brand's descriptor is read by composing the prefix with the root, so the sixteen states are generated from four positions and four reception roots rather than memorised individually. The grid itself holds exactly the sixteen on-grid states. A brand with insufficient MA or MS data sits off the grid without a label and is reported by its readable metrics with the shortfall flagged. A brand with MV = 0 is reported as AI Absent, off the grid. No brand data are plotted on this grid; it is the empty-grid case showing the taxonomy itself.



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 6: Visibility–Alignment competitive map.

MA (alignment given mention, 0–100) is on the horizontal axis and MV (visibility, 0–100) on the vertical, with presentation boundaries at MA = 50 and MV = 50. The focus brand (Acme Analytics, MV 52 / MA 67, amber Mixed band) is annotated "Acme Analytics: AI Wildcard". A second annotation, "ABC Widgets: Misrepresented AI Pariah", labels a non-focus competitor in the Visible & Misaligned quadrant. Other competitors are coloured dots without per-dot labels, and colour encodes MS band (positive green, neutral grey, mixed amber, negative red). Quadrant names appear on the teaching grid (Figure 5). Conditions for the focus brand are in the figure note and the full block in Appendix B.

9.3 Falsifiability criteria

MA incremental validity. If $\rho(\text{MV}, \text{MA}) > 0.80$ across the brand sample in a validation study under §9.6, the alignment measure adds limited diagnostic value over visibility alone. The framework's recommendation is revised accordingly. MA would be recommended only for brands that have already achieved a defined MV threshold.

MS incremental validity. If the multiple correlation R of MS regressed on (MV, MA) > 0.80 across the

brand sample, the sentiment measure adds limited diagnostic value over the other two. A single pairwise correlation is insufficient. MS could be weakly correlated with each other measure individually yet jointly predictable from both. The test addresses informational redundancy, not conditioning.

Attenuation caveat. MS carries the widest measurement error of the three metrics. Measurement error attenuates observed correlations downward, biasing the MS redundancy test toward passing. Reliability-adjusted (disattenuated) estimates would be reported alongside raw ones if a validation study were conducted.

The 3×3 correlation matrix for MV, MA, and MS would be reported as a small table from any such study. No figure.

9.4 Response sequencing by quadrant: framework-derived guidance

The quadrant position informs the appropriate response sequence. The sentiment band overlays the sequence, so in any quadrant a mixed or negative band raises the priority of sentiment work and cautions against amplifying visibility before sentiment is addressed.

Visible & Aligned (high MV, high MA): maintain visibility and alignment. Monitor MS for sentiment risks. An AI Champion holds the target state. An AI Contender (neutral reception) signals an accurate, visible brand that the engine carries without advocacy, so the work is content that earns advocacy rather than fact correction. An AI Wildcard signals divided reception over an accurate baseline, while an AI Pariah signals visible accurate description with negative reception. Both call for positioning and sentiment-driver work rather than fact correction.

Visible & Misaligned (high MV, low MA): alignment remediation before visibility investment. Building more visibility while descriptions are inaccurate amplifies representational harm. The field-level MA report identifies which facts AI gets wrong, and fixing the cited sources that carry the misaligned claims is the primary action. A Misrepresented AI Pariah requires alignment remediation before any visibility investment, since the negative reception often tracks the misrepresentation. When AI states the wrong price, the higher figure reads as "expensive" and drives the sentiment, so fixing the facts is the first move on both axes. A Misrepresented AI Wildcard sits in the same quadrant with polarised reception over the inaccurate descriptions. The same alignment-first sequence applies, with sentiment-driver work informing which misrepresentations to fix first.

Aligned but Unseen (low MV, high MA): the accuracy foundation is in place. Visibility investment can proceed with lower remediation risk. Content strategy focuses on expanding the category question space the brand appears in. An Undiscovered AI Contender signals a neutrally received, accurate brand that simply needs to be seen more often, and the visibility push can run unconditionally. An Undiscovered AI Wildcard or Undiscovered AI Pariah signals that the brand's few mentions carry polarised or adverse reception. Scaling visibility without first resolving the reception drivers amplifies the negative coverage, so sentiment-driver analysis precedes the visibility push.

Unseen & Misaligned (low MV, low MA): both dimensions require attention. Remediation and visibility work in parallel. Alignment work may have lower urgency here than in Visible & Misaligned, but is not rendered unnecessary by low MV. The framework's motivation for measuring MA separately from MV is that descriptive accuracy is a distinct property, and a brand that becomes visible while remaining inaccurate has merely converted one problem into another. Where the brand is also below the eligible-cell floor on MA and MS, the response is data first. Build sufficient mention volume that the conditional measures become reportable before sequencing remediation against them.

Standard diagnostic outputs derived from the same data (prompt gap list, per-engine comparison, field-level misalignment report, sentiment drivers, version-change decomposition) are specified in Appendix F.

9.6 Proposed validation study

The section below describes a study design, not a conducted study. The descriptions are written in the present tense for readability, but each refers to how the study would run if undertaken. The intended scope is to apply all three metrics across a minimum of 50 brands drawn from at least five commercial categories, with category stratification testing whether framework properties hold across heterogeneous competitive environments.

Design. The study uses a category benchmark set of n = 30 prompts per category, and every brand in the category is scored from the same category scan, so a single scan would score every brand against the identical stimulus (Appendix A). The engine set spans at least E = 4 engines, covering conversational, search-augmented, and the control arm specified in §7.5: Llama-3.1-8B-Instruct, or a validated substitute that passes the determinism test. The control arm's stable behaviour by construction enables noise-model validation and replicability demonstration. Its scores are excluded from all brand-level inferential tests. Each prompt-engine pair is queried k $\geq$ 3 times, with a benchmark floor of n $\times$ k $\geq$ 150 per engine for any published per-engine figures.

Power analysis. With $n = 50$ brands and the discriminant-validity threshold $\rho > 0.80$, the design has approximately 80% power for the test against materially lower true correlations at $\alpha = 0.05$ (one-tailed). Calculation: using the Fisher z-transformation, the standard error of the sample correlation is $1/\sqrt{n-3} = 1/\sqrt{47} \approx 0.146$ in z-units; the minimum detectable z-difference at 80% power and $\alpha = 0.05$ (one-tailed) is $(z_{0.95} + z_{0.80}) \times 0.146 = (1.645 + 0.842) \times 0.146 \approx 0.363$, equivalent to detecting a true $\rho \approx 0.63$ against the H_0 threshold of 0.80. The $n = 50$ design is therefore adequate for testing the $\rho > 0.80$ redundancy threshold against the alternative that the measures are usefully non-redundant in practice (true ρ in the moderate range). A more stringent contrast (e.g., true $\rho = 0.70$ vs 0.80) would require a larger sample.

Evaluator design. Two independent evaluator models from different families would score all eligible response texts used to construct eligible-cell values (per the cell-level rule in §5.3 and §6.3). A human gold standard of 200 responses, double-coded and adjudicated, would be produced for each instrument, with cell-level values constructed from the response-level scores via the §5.3 rule. The 200-response gold standard is stratified across category (5 strata, one per validation category), engine (the four commercial engines and the control arm), and claim type (Brand Profile field tier for MA: Tier 1 and Tier 2; sentiment direction for MS: positive, neutral, negative). The stratification plan would be fixed in the pre-registration. Statistical tests for both instruments follow those specified in §7.2 and §9.3. Bootstrap confidence intervals (1,000 resamples) would be reported on all published values. Because brands are nested within categories and share the category benchmark prompt set, validation analyses would report both pooled estimates and category-clustered sensitivity checks, using category-stratified bootstrap confidence intervals or mixed-effects robustness checks where appropriate.

Concurrent expert-panel criterion. As an external check on MA's convergent validity (the full argument is in §10.2), a concurrent expert panel scores brand-level representation summaries constructed from the same underlying response sample, with response excerpts blinded to the metric scores. The panel does not validate individual formulas directly. It provides a convergent-validity criterion for whether the aggregate metric profile (MV, MA, MS) corresponds to informed human judgement of brand-representation quality. The metric's agreement with the panel tests whether the framework tracks that judgement, without staking on a causal business outcome the framework explicitly disclaims (see §10.5). The panel comprises at least 5 panellists, each with at least 5 years of professional experience in marketing science, brand management, or competitive intelligence. Inter-rater agreement is reported using Krippendorff's α, with a reporting floor of $\alpha \geq 0.67$ below which the concurrent-validity evidence is treated as inconclusive. The floor sits below the instrument-validation target of $\alpha \geq 0.75$ because the panel makes holistic interpretive judgements without the codebook discipline the gold-standard coders use. The panel scores a stratified subsample of the validation brand summaries, and the result is the disattenuated correlation between panel scores and each of MV, MA, and MS, with bootstrap confidence intervals.

Pre-registration. If conducted, the study would be pre-registered prior to data collection on a public registry (OSF is the assumed venue). The pre-registration would cover the falsifiability criteria in §9.3, the gold-standard stratification plan, and the expert-panel composition and reporting floor, and would state the expected direction and magnitude of the incremental-validity tests.

Calibration goal. The primary calibration targets are the reportability conventions, namely the n $\times$ k $\geq$ 150 collection rule, the eligible-cell floor of 10, and the differentiated agreement expectations for MA and MS. Each convention would be validated or revised against observed score variance and dispersion in any conducted study. If observed score dispersion is materially below the assumed SD $\approx$ 0.25, the eligible-cell floor may be revised downward. If scores prove polarised, it should rise.

The eligible-cell floor is provisional and may differ by measure. MS is a more interpretive judgement than

MA and carries wider per-observation error, so MS may require a higher eligible-cell floor than MA to reach comparable confidence. A conducted validation study would calibrate the floor for each measure against observed score dispersion rather than assuming a single shared value.

9.7 Validity argument and current limits

The framework's validity argument states which validity properties its structure is designed to support, which the proposed validation programme (§9.6) tests, and which remain open. The taxonomy follows the unified validity framework (Messick, 1989) and the construct-validity tradition (Cronbach and Meehl, 1955; Borsboom, Mellenbergh, and van Heerden, 2004). The summary table below states each property's structural support in GBMF and its empirical status; the paragraphs that follow give the detailed argument and literature attachment.

Validity property	Structural support in GBMF	Empirical status
Content validity	Governed prompt-set construction and versioning (Appendix A)	Buyer or expert review pending validation
Construct validity	MV, MA, MS defined as separable; incremental-validity tests pre-stated (§9.3)	Empirically open
Reliability	k-run replication (§4.3); evaluator validation (§7.2); control arm (§7.5)	Empirical estimates pending
Evaluator validity	Cross-family evaluator; human gold standard; positivity-offset test (§7, Appendix D.7)	Bias measurement pending
Generalisability	Fixed-set estimand declared; prompt-set sensitivity bootstrap (§4.3)	Bounded by declared conditions
Criterion / convergent validity	Tier 1 minimum (§5.2); concurrent expert panel (§9.6); adjacent factuality/sentiment literatures	Convergence tests to be conducted
Consequential validity	Governance constraints and manipulation-risk monitoring (§§8.5, 8.6, 10.4)	Requires field monitoring

Content validity concerns whether the declared prompt set represents the category's buyer question space. The framework treats representativeness as a governance property under Appendix A.2 rather than an empirical property of the metric itself, and the proposed validation programme adds buyer or expert review of prompt naturalness and category coverage. The structural argument is in place, and the empirical demonstration is open pending validation.

Construct validity concerns whether MV, MA, and MS measure the three separable properties the framework claims, namely presence, factual conformance given mention, and brand-directed sentiment given mention. The structural commitment is to construct separability, declared explicitly and distinguished from statistical independence (§9.1). The incremental-validity tests in §9.3 are the construct-validity tests for non-redundancy, and the concurrent expert-panel criterion in §9.6 is the convergent-validity test. Structurally committed, empirically open.

Reliability concerns whether scores are stable under repeated measurement when the underlying phenomenon has not changed. The k-run replication (§4.3) and the control arm (§7.5) jointly address reliability. The former bounds within-scan stochastic noise, while the latter isolates protocol noise from engine signal. Across-scan test-retest reliability and inter-evaluator reliability (Krippendorff's α targets at §7.2) are part of the proposed validation programme. Structurally addressed, empirical estimates open.

Evaluator validity concerns whether the LLM evaluators apply the rubric without introducing systematic bias. Cross-family evaluator selection (§7) reduces shared pre-training bias, and the human gold-standard calibration step (§7.1, §7.2) defines the agreement targets the evaluator must meet. Known evaluator risks including position bias, verbosity bias, self-preference, and prompt sensitivity are documented in the LLM-as-judge bias literature (Zheng et al., 2023), and the MS positivity-offset test (Appendix D.7) is one explicit bias check the production protocol runs. Structurally addressed with documented risks, and empirical bias measurement is part of the proposed programme.

Generalisability concerns whether scores generalise across categories, brands, time periods, and prompt-set instances. The fixed-set estimand declines to claim generalisation across prompt-set instances, and the prompt-set sensitivity bootstrap (§4.3) quantifies how much a headline shifts under comparable alternative prompt selections. Across categories and brands, the proposed validation programme uses category-clustered sensitivity checks (§9.6). Across time, the version-marked series and the no-adjustment principle (§8.4) preserve longitudinal integrity. Scope is declared narrowly, and generalisability is bounded by the declared conditions in every published score.

Criterion and convergent validity concern whether scores agree with external benchmarks that should track the same underlying constructs. For MV, the source-level GEO measurement reviewed in §2.1 is the closest adjacent benchmark; criterion-validity testing against a comparable brand-entity benchmark awaits the development of one. For MA, the FactScore and SAFE literature (§2.2) provides the closest external benchmark, and convergence between MA on Tier 1 fields and external-corpus grounding is one criterion-validity test. For MS, brand-directed sentiment can be compared against human expert panels (§9.6 concurrent expert-panel criterion). Criterion tests are specified in §9.6, and results are open.

Consequential validity concerns whether use of the metric produces the consequences the framework supports without producing unintended harms. The principal known risk is content-ecosystem manipulation (§10.4). Once the metric becomes commercially used, brand actors may optimise the answer space rather than the brand itself. The framework's governance commitments (declared prompt sets, cross-family evaluators, control arm, version transparency) constrain the gaming surface, but no governance arrangement eliminates it. Independent governance of prompt sets and engine lists, and a published gaming-vector watch, are required as the instrument's use scales. Governance constraints are specified, and consequential-validity monitoring is part of the proposed programme rather than a property of the formulas.

Validity status summary. The framework presently makes a structural validity argument. Each construct has a definition, a measurement procedure, and a place in the integrated specification. The empirical validity argument, that scores reliably track the intended constructs, generalise across the declared scope, and produce the intended consequences without harms, depends on validation evidence the paper does not yet provide. The validation programme in §9.6 is designed to produce that evidence.

---

10. Discussion

10.1 Scope and limitations of MV

MV measures whether a brand is mentioned, not how. Response position, recommendation framing, and the character of the mention are not captured in the mention rate. They are measurable as secondary diagnostic signals alongside MV, not inside it.

MV as specified is a prompt-uniform measure. Every prompt in the declared set contributes equally to the headline score regardless of real-world query frequency. Prompt-uniform MV and query-volume-weighted brand exposure are different constructs. A brand with MV 52 on a researcher-constructed prompt set may have materially different effective exposure if the prompts driving the highest AI query traffic are not proportionally represented in the sample. Query-volume weighting would require dependency on external traffic data, which changes over time and introduces a confound that the current framework deliberately avoids. The prompt set's representativeness of real query distributions is addressed in governance (Appendix A) rather than in the formula.

Reliability concerns specific to longitudinal tracking (within-scan stochasticity, the k-run sampling variance redrawn each scan, and engine drift between scans) are addressed by the uncertainty decomposition in §4.3

and the control arm in §7.5. A formal across-scan reliability study on a stable commercial system is part of the proposed validation design.

Multilingual instability. MV is specified for a single measurement language. Brand representations differ substantially across languages. Cross-lingual measurement requires separate prompt sets, separate brand profiles, and separate computations. Multilingual extension is not addressed here.

Retrieval volatility. For search-augmented engines, mention patterns may shift with indexed content independently of model weight changes. A brand's MV may decrease following negative press coverage because retrieved documents changed, not because latent representations changed. The version-marked series makes such shifts detectable and dateable.

10.2 Scope and limitations of MA

The MA depends on Brand Profile quality. A sparse or incorrectly specified Brand Profile will fail to detect real misalignments or misclassify correct descriptions as misaligned. Brand Profile governance is as consequential as prompt set governance.

MA scoring depends on the evaluator model. Different models may apply the scoring rubric differently on graded fields requiring interpretation. Evaluator model version pinning and documentation in all published reporting address but do not eliminate this dependency. LLM evaluators introduce known risks. Evaluator drift as model versions change, and systematic agreement inflation when the evaluator shares training data with the evaluated engine, are the principal ones. Cross-family evaluator selection reduces the second risk. Human-calibrated validation addresses the first.

The MA's eligible-cell coverage statistic is a first-order diagnostic. Consistent low coverage means the brand is named but not described. AI systems include it in lists but provide no factual characterisation. This state has a distinct remediation path, different from both low visibility and low alignment.

Convergent validity. MA measures conformance to the brand's own declared facts (§5.2). That estimand is auditable across time because the Brand Profile is versioned and held fixed, and the measure has no built-in anchor to external truth. The Tier 1 minimum (§5.2) requires at least three externally verifiable fields in any Brand Profile used for benchmark MA publication. Those fields are the framework's structural floor against the case where a brand curates a profile that simply matches what AI already says. Tier 1 fields do not equate to full external-corpus grounding (see §2.2 on FactScore and SAFE). The residual gap is what the concurrent expert-panel criterion in §9.6 is designed to test. Whether informed human judgement of brand-representation quality agrees with the metric is evaluated against a stated inter-rater agreement reporting threshold. Three elements jointly establish MA's convergent validity: the Tier 1 minimum (a structural floor in the specification), the concurrent expert panel (an empirical test in the validation study), and the engagement with the LLM factuality-evaluation literature on the choice of ground truth (a methodological positioning). None alone is sufficient.

10.3 Scope and limitations of MS

MS is the framework's most interpretive measure. Sentiment classification is less determinate than factual alignment checking. Even human annotators disagree more on sentiment than on factual correspondence. The expected inter-rater reliability ceiling for MS is below MA's, and this paper states differentiated expected agreement levels for the two measures rather than presenting them as equally precise.

The brand-directed definition in Appendix D is what prevents silent divergence between implementations. An evaluator that scores text polarity rather than brand-directed sentiment will systematically mis-score category-critical-but-brand-positive responses, producing non-comparable scores across implementations that differ on this point.

Aspect-level sentiment disaggregation is noted as a diagnostic extension but is not specified here.

10.4 Adversarial optimisation

As any measurement framework attracts commercial attention, it becomes a target for optimisation pressure. The governance-dependent resistance discussed in §8.6 addresses prompt cherry-picking and Brand Profile

gaming. Content-ecosystem manipulation sits beyond that boundary: a brand can manufacture mention frequency by flooding the web with synthetic third-party endorsements, and if those are indexed and trained on, MV rises and the movement is real. The monitoring signal is sustained score inflation without corresponding change in independently verifiable brand standing.

10.5 Implications for research

The three-measure framework opens research programmes currently constrained by measurement heterogeneity. Cross-category studies can now separate visibility effects from alignment and sentiment effects. Longitudinal tracking can test whether corrective brand actions improve MA and MS over time. Whether AI mention rates have measurable relationships to commercial outcomes at the brand level (purchase intent, consideration, revenue) remains an open empirical question that this framework makes addressable but does not resolve.

Downstream-outcome validation (consideration, purchase intent, brand-attributed traffic, revenue) sits outside the scope of this paper and is not part of the proposed validation design. The concurrent expert-panel criterion in §9.6 does not substitute for it. The panel tests whether the metrics track informed human judgement of brand-representation quality at the level of the AI response, while downstream-outcome studies test whether brand-representation quality, as measured here, in turn predicts commercial behaviour. Downstream outcomes are confounded by non-AI determinants of consumer behaviour at this stage of the technology's diffusion, so a weak observed correlation would unfairly indict a well-functioning measurement instrument. The framework's data make such studies possible if anyone undertakes them. This paper makes no commitment to that work.

---

11. Conclusion

The Generative Brand Mention Framework specifies three separable measures of brand representation in AI-generated answers (visibility, alignment, and sentiment) under declared prompt sets, declared engine sets, declared evaluator architecture, and pre-stated falsifiability thresholds. The full specification is published openly so that any measurement provider can implement, replicate, contest, or extend the design. The thresholds at which the framework would be revised are stated in advance, ahead of any validation study that tests them.

The validation design in §9.6 specifies how the empirical redundancy of MA and MS with MV would be measured across a brand sample. If a conditional measure is largely predictable from the others ($\rho > 0.80$ in the incremental-validity tests), its incremental diagnostic value is limited and the framework's recommendations would revise accordingly. The separability of the constructs is a property of the framework's design, and the strength of their measured relationship is a property of the field.

Implementations, replications, and disagreements about the calibration thresholds or convergent-validity assumptions are welcomed, and would inform any subsequent versions of the specification.

Data Availability

The Brand Profile specification, the MA scoring rubric, the MS codebook, and the prompt generation methodology are published as open specifications in the appendices of this paper. The full per-prompt per-engine mention rates for the Appendix B worked example are published as supplementary data in `gbmf-worked-example-mention-rates.csv` (30 prompts × 5 engines), so the headline MV, per-engine MV, reach, intensity, and HHI values in Appendix B reproduce from a single open dataset. Aiviara Research maintains a reference implementation. Independent implementations are encouraged, and all methodological inputs required for independent implementation are published here.

Acknowledgements

The author thanks early reviewers whose feedback improved this draft.

Disclosure on AI-Assisted Preparation

The author used AI-assisted drafting and critique tools during preparation of this manuscript. The author reviewed, edited, verified, and takes full responsibility for all claims, references, formulae, and conclusions.

Competing Interests

The author is a founder of Aiviara Research, which maintains the reference implementation of the framework described in this paper. The framework is published as an open methodology, and the commercial interest is in implementation and application, not in restricting independent use.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD '24), 5–16. ACM. https://doi.org/10.1145/3637528.3671900

Ahrefs. (2025a). New study: AI assistants prefer to cite 'fresher' content (17 million citations analyzed). *Ahrefs Blog*. https://ahrefs.com/blog/do-ai-assistants-prefer-to-cite-fresh-content/

Ahrefs. (2025b). 67% of ChatGPT's top 1,000 citations are off-limits to marketers (+ more findings). *Ahrefs Blog*. https://ahrefs.com/blog/chatgpts-most-cited-pages/

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '04), 25–32. https://doi.org/10.1145/1008992.1009000

Chen, M., Wang, X., Chen, K., & Koudas, N. (2025). Generative Engine Optimization: How to dominate AI search. arXiv:2509.08919 [cs.IR]. [Preprint; not peer-reviewed at time of writing.]

Conductor. (2026). *The 2026 AEO / GEO Benchmarks Report*. Conductor Research. https://conductor.com/academy/aeo-geo-benchmarks-report/

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. https://doi.org/10.1037/h0040957

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/

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *PNAS*, 120(30). https://doi.org/10.1073/pnas.2305016120

Jackman, S. (2005). Pooling the polls over an election campaign. *Australian Journal of Political Science, 40*(4), 499–517.

Keller, K. L. (1993). Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing, 57*(1), 1–22.

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

Kumar, A., & Palkhouski, L. (2025). AI answer engine citation behavior: An empirical analysis of the GEO16 Framework. arXiv:2509.10762 [cs.AI]. [Preprint; not peer-reviewed at time of writing.]

Lavidge, R. J., & Steiner, G. A. (1961). A model for predictive measurements of advertising effectiveness. *Journal of Marketing, 24*(6), 59–62.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

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* (EMNLP 2023), 12076–12100. https://aclanthology.org/2023.emnlp-main.741/

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2016). SemEval-2016 task 6: Detecting stance in tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), 31–41. https://doi.org/10.18653/v1/S16-1003

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology*, 17(3), Article 26. https://doi.org/10.1145/3003433

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1), 3–26. https://doi.org/10.1075/li.30.1.03nad

OtterlyAI. (2026). Llms.txt experiment: What marketers get wrong about llms.txt. *OtterlyAI Blog*. https://otterly.ai/blog/the-llms-txt-experiment/

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, 81(12), 46–54.

Rhoades, S. A. (1993). The Herfindahl-Hirschman Index. *Federal Reserve Bulletin*, 79, 188–189.

Rogan, W. J., & Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology, 107*(1), 71–76.

Sakai, T. (2014). Statistical reform in information retrieval? *SIGIR Forum*, 48(1), 3–12. https://doi.org/10.1145/2641383.2641385

Schulte, J., Bleeker, M., & Kaufmann, P. (2026). Don't measure once: Measuring visibility in AI search (GEO). arXiv:2604.07585 [cs.IR]. [Preprint; not peer-reviewed at time of writing.]

Voorhees, E. M., & Harman, D. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.

Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv:2304.04339 [cs.CL]. [Preprint; not peer-reviewed at time of writing.]

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., & Le, Q. V. (2024). Long-form factuality in large language models. *Advances in Neural Information Processing Systems* 37 (NeurIPS 2024). arXiv:2403.18802.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Xing, E. P., Gonzalez, J. E., Stoica, I., & Hashimoto, T. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36 (NeurIPS 2023): Datasets and Benchmarks Track. arXiv:2306.05685.

Appendix A: Prompt Generation and Governance

A.1 Generation process

Two prompt-set types are maintained and never merged:

Prompt-set type	Used for	Contains focal-brand name	Version identifier	Comparable across brands
Category benchmark set	MV/MA/MS benchmark figures and the competitive map	No	Category-level	Yes
Brand diagnostic set	Single-brand diagnostics, named comparisons, remediation	May	Brand-level	No

Only the category benchmark set feeds published benchmark figures and the quadrant map. A single category scan therefore scores every brand in the category against the identical stimulus, which is what makes the competitive map valid. Brand diagnostic sets are maintained separately and labelled non-comparable to benchmark figures.

For the **category benchmark set**, the prompt generation process takes as input (i) category classification; (ii) the competitor/brand set defining the category (used to scope the category, not to seed focal-brand prompts); (iii) target use cases describing how the category's buyers use the product type. The benchmark set must not contain any brand name from the measured competitive set, and brand names enter only diagnostic sets.

A primary agent generates candidate prompts across four query types: (1) categorical ("what are the best X tools for Y?"), (2) comparative ("how do leading platforms in this category compare?"), (3) use-case ("I need a tool for doing Z, what do you recommend?"), (4) problem-solution ("I have problem P with my current setup, what should I look for?"). A secondary agent reviews for coverage balance (at least 20% of prompts in each type), naturalness, category specificity, and phrasing diversity (no more than 20% sharing the same sentence structure).

After quality review, 30–50 prompts are selected, assigned a category-level version identifier in the format `[CategoryID]-v[YYYY-MM-DD]`, and stored as an immutable record.

For the **brand diagnostic set**, generation may take a brand name and competitor list as input, and prompts may name the focal brand. The diagnostic set is assigned a brand-level version identifier in the format `[BrandID]-v[YYYY-MM-DD]` and is held separately from the benchmark set. Diagnostic-set scores are labelled non-comparable to MV/MA/MS benchmark figures.

The prompt set size bounds are a provisional convention. $n = 30$ prompts at $k = 5$ runs meets the $n \times k \geq 150$ collection requirement per engine, while $n = 50$ at $k = 3$ also meets it. Both are acceptable.

A.2 Governance rules

Subsequent measurement periods use the identical prompt set. A prompt set may only be updated when there is a documented change in the category's scope, or a previously relevant category question type has become obsolete. Each update requires a new version identifier, a logged break in the time series, and the overlap-window transition protocol described in §8.4.

The governance rationale is direct. The prompt set defines what is being measured. An operator who updates the prompt set in response to brand performance invalidates the comparative integrity of all historical scores. The canonical set must be held fixed by a party whose incentive is measurement integrity, not brand performance.

A.3 Mention detection

A brand is mentioned if the response contains the brand's canonical name or a recognised abbreviation. Detection is case-insensitive. Partial matches are logged separately and excluded from the primary binary mention outcome. This detection rule is held constant across all measurement periods.

A.3.1 Entity-resolution protocol Brand entity resolution is a documented hard problem in named entity recognition (Nadeau and Sekine, 2007). The framework requires a documented protocol per implementation, with the following minimum rules.

Canonical name and aliases. Each brand has one canonical name (the form used in the Brand Profile) and a published alias list covering registered marks, common abbreviations, and stylised forms. Alias additions and removals are version-controlled with the Brand Profile.

Company brand and product brand. Where a brand has both a company-level identity and one or more product-line identities, each is treated as a separate measured brand by default. Aggregation under a single canonical identity is permitted only when the Brand Profile declares them as a single measured identity and documents the rule. Mixed measurement (sometimes aggregated, sometimes separate) is not permitted within a single scan window.

Subsidiaries and acquired brands. Subsidiaries and acquired brands are measured separately by default. Aggregation under a parent identity is permitted only when the parent's Brand Profile enumerates the included sub-identities and their aggregation rule, with the decision logged in the reporting conditions block.

Ambiguous abbreviations. Abbreviations that match multiple entities require disambiguation context. The default is to exclude bare-abbreviation matches and count only abbreviations that appear in disambiguating context, where the abbreviation is co-located with the canonical name in the same response or with a documented disambiguating cue. Pure-abbreviation matches in non-disambiguating context are logged separately under a flagged ambiguous-mention counter that does not contribute to the headline MV.

Rebrands. When a brand changes name during a measurement period, the old and new names are both counted as mentions of the post-rebrand entity for the transition window. The Brand Profile records the rebrand event date and the alias retirement schedule. After the retirement window, the old name is excluded.

Pluralisations and translations. Standard English pluralisations are included by default. Translated forms are excluded from a canonical English-language scan unless explicitly listed in the alias set. Multilingual measurement uses separate language-specific prompt sets and Brand Profiles (§10.1).

Implicit references. URLs, citation markers, or product images that imply a brand without naming it are excluded from MV. MV is a mention measure, and implied or visual brand presence is out of scope for this specification.

Near-name variants. Variants close to but distinct from the canonical form or any registered alias, for example misspellings or near-homophones, are logged separately under a near-mention counter and do not contribute to the headline MV.

Each rule is implemented as a documented decision in the implementation's detection configuration, with audit logs accessible for independent review. Implementers report their canonicalisation decisions as part of the reporting conditions block. Entity-resolution cases outside this protocol (compound identities, joint ventures, white-label or OEM relationships) require implementation-specific decisions that are not modelled here.

A.4 Version-transition protocol

At a prompt-set version change, both versions run concurrently for one overlap scan window. The published output at the boundary includes three quantities:

1. The old-set MV (the score that continues the old series)
2. The new-set MV (the score that opens the new series)
3. A decomposition: how much of the difference is attributable to movement on prompts present in both versions versus position in prompts present only in the new version

The series then continues with new-set scores. No splicing or rescaling occurs. The old-set history stands as measured, and the new series starts from the new-set score at the boundary.

Appendix B: Worked Example, End-to-End

Setup:

- Brand: Acme Analytics (synthetic)
- Prompt set (category benchmark): n = 30 prompts, version B2B-ANALYTICS-v2025-11-01 (Acme's worked-example scan; the CategoryID names the analytics category Acme competes in, not Acme itself, so the same prompt set scores every brand in the category; see Appendix A.1)
- Engines: E = 5 (ChatGPT [GPT], Claude [CLD], Perplexity [PPX], Gemini [GEM], Microsoft Copilot [COP])
- Runs: k = 5 per prompt-engine pair; n × k = 150 per engine, meeting the benchmark publication threshold
- Measurement period: 2026-06

Acme Analytics Brand Profile (ACME-v2025-11-01):

Field	Value	Type
Primary category	B2B analytics and reporting software	Graded
Pricing entry point	$89/month	Thresholded
Key claim 1	Integrates with all major CRM platforms	Binary
Key claim 2	No-code dashboard builder	Binary
Disambiguation	Acme Analytics is not a CRM; it does not store contact data	Graded
Founding year	2019	Thresholded
Company type	Private	Binary

Part 1: MV computation

For each prompt p and engine e, $r_e(p)$ is the fraction of k = 5 runs in which the engine mentioned Acme Analytics.

Per-engine mention rates across 30 prompts (showing representative rows; remaining prompts follow a declining pattern):

Prompt	GPT r(p)	CLD r(p)	PPX r(p)	GEM r(p)	COP r(p)
p1	1.00	1.00	1.00	1.00	1.00
p2	1.00	1.00	0.80	1.00	0.80
p3	1.00	0.80	1.00	0.80	0.60
p4	0.80	1.00	0.80	0.80	0.60
p5	0.80	0.80	0.60	0.80	0.40
p6	0.80	0.80	0.60	0.60	0.40
p7	0.60	0.60	0.40	0.60	0.40
p8	0.60	0.60	0.40	0.60	0.20
p9	0.60	0.40	0.20	0.40	0.20

Prompt	GPT r(p)	CLD r(p)	PPX r(p)	GEM r(p)	COP r(p)
p10	0.40	0.40	0.20	0.40	0.20
p11–p30	See distribution below	See distribution below	See distribution below	See distribution below	See distribution below

p11–p30 mention rate distributions (# prompts at each rate level):

Engine	1.0	0.8	0.6	0.4	0.2	0.0
GPT	3	4	6	3	0	4
CLD	3	4	5	2	0	6
PPX	3	4	2	1	0	10
GEM	2	5	4	2	0	7
COP	1	3	5	2	0	9

All p11–p30 rows sum to 20 prompts. Non-zero count per engine: GPT 16, CLD 14, PPX 10, GEM 13, COP 11. The four prompts at which GPT has rate 0.0 are unmentioned by all engines, so reach is determined by GPT's non-zero set: 26 of 30 prompts (87%).

Derivation check. $MV_e = 100 \times (1/30) \times \sum_p r_e(p)$. Rows p1–p10 contribute: GPT 7.6, CLD 7.4, PPX 6.0, GEM 7.0, COP 4.8. Rows p11–p30 contribute the remainder: GPT 11.0, CLD 10.0, PPX 7.8, GEM 9.2, COP 7.2. Totals: GPT 18.6, CLD 17.4, PPX 13.8, GEM 16.2, COP 12.0. Mentioning cells (rate > 0 on any run): p1–p10 contribute 10 non-zero prompts per engine (50 cells); p11–p30 contribute GPT 16, CLD 14, PPX 10, GEM 13, COP 11 (64 cells). Total mentioning cells: 26 GPT + 24 CLD + 20 PPX + 23 GEM + 21 COP = **114**.

Per-engine MV (mean mention rate × 100, across all 30 prompts):

Engine	MV_e
GPT	62
CLD	58
PPX	46
GEM	54
COP	40

Headline MV = (62 + 58 + 46 + 54 + 40) / 5 = 260 / 5 = **52**

Reach: 87% of prompts (26 of 30 had at least one mention from at least one engine). Per-engine reach: GPT 87%, CLD 80%, PPX 67%, GEM 77%, COP 70%.

Per-engine intensity (each engine's mean mention rate among the prompts that engine reaches): GPT 0.72, CLD 0.73, PPX 0.69, GEM 0.70, COP 0.57. Headline intensity, computed as the equal-weighted mean of per-engine intensities (the same aggregation used for headline MV in §4.2): (0.72 + 0.73 + 0.69 + 0.70 + 0.57) / 5 = **0.68**. Cross-engine spread is substantial (range 0.57–0.73), and the per-engine values are reported alongside the headline rather than absorbed into it.

Concentration (HHI over prompt contribution shares): **0.047** (uniform baseline for 30 prompts: 0.033; approximately 1.4× uniform, indicating broadly distributed visibility close to the uniform baseline).

HHI is derivable from the full per-prompt per-engine mention rates published as supplementary data (`gbmf-worked-example-mention-rates.csv`). The table above lists representative rows, and the distribution table at p11–p30 summarises the rest in aggregate form.

Part 2: MA computation

Across 114 mentioning cells in total (26 GPT + 24 CLD + 20 PPX + 23 GEM + 21 COP), 70 were eligible (triggering ≥ 2 Brand Profile fields). MA coverage = 70 / 114 = **61%**.

Sample alignment evaluation, p1 (GPT response):

> "Acme Analytics is a B2B analytics platform offering no-code dashboards and CRM integrations, starting from around $120/month."

Field	Triggered	AI claim	Brand Profile	Score
Primary category	Yes	"B2B analytics platform"	"B2B analytics and reporting software"	1.0
Pricing entry point	Yes	"$120/month"	"$89/month"	0.0 (35% above tolerance)
CRM integration	Yes	"CRM integrations"	Confirmed	1.0
No-code dashboard	Yes	"no-code dashboards"	Confirmed	1.0
Disambiguation	Not triggered	N/A	N/A	excluded
Founding year	Not triggered	N/A	N/A	excluded
Company type	Not triggered	N/A	N/A	excluded

$A(r, F_B) = (1.0 + 0.0 + 1.0 + 1.0)/4 = \mathbf{0.75}$

Sample alignment evaluation, p2 (CLD response):

> "Acme Analytics provides no-code reporting tools for B2B teams, with plans starting at $89/month. It connects to major CRMs and was founded in 2019."

Field	Triggered	AI claim	Brand Profile	Score
Primary category	Yes	"no-code reporting tools for B2B teams"	"B2B analytics and reporting software"	0.5
Pricing entry point	Yes	"$89/month"	"$89/month"	1.0
CRM integration	Yes	"major CRMs"	Confirmed	1.0
No-code dashboard	Yes	"no-code reporting"	"no-code dashboard builder"	0.5
Founding year	Yes	"2019"	"2019"	1.0

$A(r, F_B) = (0.5 + 1.0 + 1.0 + 0.5 + 1.0)/5 = \mathbf{0.80}$

Aggregated per-engine MA:

Engine	Eligible cells	Per-engine MA
GPT	16	71
CLD	15	76
PPX	12	63
GEM	16	68
COP	11	59

All five engines clear the 10-cell floor.

Headline MA = (71 + 76 + 63 + 68 + 59) / 5 = 337 / 5 = **67**

Field-level misalignment: pricing is the dominant misalignment. Approximately half of all eligible mentioning cells included a pricing claim outside the ±15% tolerance. Category and feature claims are well-aligned.

Part 3: MS computation

70 eligible mentioning cells assessed for brand-directed sentiment. (The MA and MS eligible cell counts coincide at 70 in this illustration. This is a property of the synthetic dataset, not a general feature. The two eligibility criteria may yield different counts in practice, as explained in Appendix D.4.)

Engine	Eligible cells	Positive	Neutral	Negative	Net
GPT	16	8 (50%)	4 (25%)	4 (25%)	+25
CLD	15	9 (60%)	3 (20%)	3 (20%)	+40
PPX	12	5 (42%)	3 (25%)	4 (33%)	+9
GEM	16	8 (50%)	4 (25%)	4 (25%)	+25
COP	11	6 (55%)	2 (18%)	3 (27%)	+28

Headline shares (equal-weighted mean of per-engine shares): 51% positive / 23% neutral / 26% negative. Headline net (mean of per-engine nets) = (25 + 40 + 9 + 25 + 28) / 5 = **+25**.

Display band: Mixed (positive 51% $\geq$ 25% and negative 26% $\geq$ 25%).

Sentiment drivers: positive characterisations highlight integration breadth and ease of use. Negative characterisations focus on pricing uncertainty (AI responses stating the wrong price, which then read as "expensive relative to alternatives").

From scores to descriptor

A brand's three scores resolve to one brand-representation state by two lookups, applied only when the brand clears the eligible-cell floor on both conditional measures. Position fixes the prefix. MV and MA are each read against the 50 boundary to place the brand in a quadrant, which sets the prefix (none / Undiscovered / Misrepresented / Misrepresented and Undiscovered). Sentiment fixes the root. The MS display band sets the reception root (AI Champion / AI Contender / AI Wildcard / AI Pariah). The descriptor is prefix plus root. A brand with insufficient MA or MS data takes no state label and is reported by its readable metrics with the shortfall flagged, in any quadrant. A brand with MV = 0 is reported as AI Absent.

For Acme Analytics, MV 52 is at or above 50 (visible) and MA 67 is at or above 50 (aligned), placing Acme in Visible & Aligned, so the prefix is none. The MS shares are 51% positive / 23% neutral / 26% negative. Positive and negative both clear 25%, so the band is Mixed, and the root is AI Wildcard. Acme Analytics is therefore an **AI Wildcard**. The absent prefix marks a visible and accurately described brand. The only open question is reception, and Acme's is divided, praised for capability and criticised on price.

Reporting conditions block

- Engine list: CLD-COP-GEM-GPT-PPX (non-standard subset of the standard list; AIO omitted for collection practicality in this illustration)
- Prompt set: B2B-ANALYTICS-v2025-11-01, n = 30
- Measurement period: 2026-06
- k = 5 runs per prompt-engine pair (n × k = 150 per engine; meets benchmark publication threshold)
- Evaluator: claude-opus-4-6 (cross-family for GPT, GEM, COP responses; same-family caveat noted for CLD responses per §7.3)
- Eligible MA cells: 70; coverage = 61%
- Eligible MS cells: 70

Summary and primary-action reading

Metric	Value	Notes
MV	52	87% reach (per-engine 67–87%); headline intensity 0.68 (per-engine 0.57–0.73, equal-weighted mean); HHI = 0.047 vs uniform baseline 0.033 (1.4× uniform); derivable from `gbmf-worked-example-mention-rates.csv`
MA	67	61% coverage (39% bare-name mentions); pricing is dominant misalignment
MS net	+25	Mixed band (51%/23%/26%); pricing uncertainty drives negative characterisations
State	**AI Wildcard**	Visible & Aligned (no prefix) + Mixed band; reception is split, factual baseline is intact

Plain-language interpretation: Acme Analytics appears in just over half of category-relevant AI responses across the declared engines. When it appears, the average eligible mentioning cell receives an alignment score of 67/100 against the Brand Profile. When alignment fails, pricing is almost always the reason. Sentiment is mixed. The brand is praised for its product capabilities but criticised in contexts where AI states an incorrect price, making it appear expensive relative to the accurate figure. The primary action is correcting pricing claims in AI-cited third-party sources before pursuing further visibility growth. Because the negative sentiment tracks the pricing misalignment, fixing the facts is likely to improve both MA and MS simultaneously.

Appendix C: MA Scoring Rubric

This appendix constitutes the published scoring rubric for MA. Third-party implementations of MA must use this rubric or explicitly document deviations.

C.1 Field types and default classification

Brand Profile fields are classified as:

- **Binary fields:** Company type. Score 0 (incorrect) or 1 (correct).
- **Thresholded fields:** Pricing entry point, founding year. Score 0, 0.5 (partially correct within stated tolerance), or 1 (correct).
- **Graded fields:** Primary category, key differentiating claims, disambiguation statements. Score 0 (wrong), 0.5 (partially correct), or 1 (correct).

Tier 3 positioning claims are excluded from Brand Profile fields for MA scoring. Only Tier 1 and Tier 2 fields are evaluated.

C.2 Grade definitions

Primary category:

- 1.0: Category stated and matches Brand Profile within normal vocabulary variation ("analytics and reporting platform" matches "B2B analytics and reporting software"; "SaaS tool" does not match "analytics platform")
- 0.5: Partially correct, correct domain but missing a key qualifier, or uses a broader category without contradiction
- 0.0: Wrong category, or directly contradicts a disambiguation statement

Key differentiating claims:

- 1.0: Claim stated and factually correct
- 0.5: Claim implied but not explicitly stated, or stated with a minor misalignment that does not fundamentally change the meaning
- 0.0: Claim explicitly contradicted, or a directly opposite claim made

Disambiguation statements:

- 1.0: Response makes no claim that the disambiguation statement prohibits
- 0.5: Response makes a claim that partially overlaps with a prohibited category (borderline case)
- 0.0: Response makes a claim explicitly prohibited by the disambiguation

Pricing entry point (thresholded):

- 1.0: Stated price within ±5% of Brand Profile value
- 0.5: Stated price within ±15% of Brand Profile value
- 0.0: Stated price outside ±15%, directly contradicted, or a specific price is claimed where none exists

Founding year (thresholded):

- 1.0: Exact match
- 0.5: Within ±2 years (AI training data may conflate founding with funding rounds or product launches)
- 0.0: More than ±2 years discrepancy, or directly contradicted

Company type (binary):

- 1.0: Correct primary classification (private, public, acquired, etc.)
- 0.0: Wrong primary classification

C.3 Non-triggered field handling

A field is triggered if the response contains claims that can be evaluated against it. Non-triggered fields are excluded from the computation for that response.

$$A(r, F_B) = \frac{\sum_{f \in T} \text{score}(f, r, F_B)}{|T|}$$

where T is the set of triggered Brand Profile fields in response r. Silence about a field (not mentioning pricing) is treated differently from stating a wrong value (triggered and scored 0). Absence of information is not misalignment.

C.4 Minimum triggered-field threshold

Responses triggering fewer than two fields are excluded from MA computation. A response mentioning the brand name only incidentally, without descriptive claims, provides no information for alignment evaluation.

C.5 Field weighting

All triggered fields contribute equally to $A(r, F_B)$ in the default rubric. Implementations applying differential field weights must document the weighting scheme, and their MA scores are not comparable to equal-weight scores under this specification.

C.6 Evaluator prompt template

The evaluator receives:

1. The Brand Profile with all fields and current values
2. The AI response text to evaluate
3. The instruction: "For each Brand Profile field that is directly addressed in this response, assign a score according to the rubric below. Do not assign scores for fields not addressed. Output JSON only: `{"field_name": score, "triggered_fields": ["field1", ...], "notes": "..."}`. Do not include fields not triggered."
4. The full scoring rubric from Appendix C §C.2, appended to the prompt

The evaluator session must not have prior exposure to the responses being evaluated within the same session. Data collection and evaluation must be conducted in separate sessions.

Appendix D: MS Sentiment Codebook

D.1 Construct definition

MS measures **sentiment toward the brand**, namely whether the AI response speaks positively or negatively about the brand, not the overall tone of the text in which the brand appears.

The unit of analysis is the brand-directed characterisation. 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 a category but identifies the brand as a poor fit for certain users is **negative** for the brand.

This definition is operationally distinct from text-level sentiment classification. An evaluator classifying text polarity rather than brand-directed sentiment will systematically mis-score category-critical-but-brand-positive responses, producing non-comparable results. The worked counter-example below illustrates the critical case.

The closest technical literatures are target-dependent sentiment analysis and stance detection. Both evaluate sentiment directed at a specified target rather than inferred from whole-text polarity (Mohammad et al., 2016; Mohammad, Sobhani, & Kiritchenko, 2017). The 2017 work extends the SemEval-2016 task by jointly annotating stance and sentiment in the same tweet sample, showing that the two are correlated but distinguishable. MS is target-dependent (brand-directed) sentiment, and stance detection is an adjacent methodology the codebook draws on, not a synonym for the construct. These references are included for implementer information, and the term "stance" does not appear in reader-facing sections of the paper.

D.2 Classification categories

- **Positive:** the response expresses approval, endorsement, recommendation, or favourable comparison toward the brand, in the context of the claimed attribute or the query's use case.
- **Neutral:** the response mentions the brand without expressing valenced characterisation, or provides a balanced characterisation with no net directional evaluation.
- **Negative:** the response expresses disapproval, criticism, unfavourable comparison, or a warning about the brand in the context of the claimed attribute or the query's use case.

When a response contains both positive and negative characterisations of the brand, the dominant direction determines the classification. If no dominant direction is clear, classify as neutral with a note for human review.

D.3 Worked counter-example (critical case)

Response text: > "Most analytics tools in this category are bloated and overpriced for small teams. Acme Analytics is the exception: it keeps the interface lean and the entry price is reasonable."

Text-level sentiment: negative (the text criticises "most tools").

Brand-directed sentiment: positive (Acme Analytics is specifically praised, distinguished from the negatively described category norm).

Correct MS classification: Positive.

Evaluators trained on text-level sentiment will mis-score this case as negative. This is the single most common failure mode in brand-directed sentiment evaluation. The instruction to the evaluator must explicitly state that the evaluator should classify the sentiment expressed toward the named brand, not the sentiment of the passage overall.

D.4 Eligibility for MS evaluation

A mentioning cell is eligible for MS if at least one mentioning run in the cell contains a substantive characterisation of the brand beyond bare-name inclusion. A response listing the brand in a set ("top tools include X, Y, Acme, and Z") without any evaluative content is not eligible for MS. If no run in a mentioning cell contains an eligible characterisation, the cell counts as a mentioning but MS-ineligible cell and is excluded from the MS denominator (see §5.3 and §6.3 for the cell-level rule).

The MS eligible-cell set may differ from the MA eligible-cell set. MS requires at least one eligible brand-directed evaluative characterisation within the mentioning cell, while MA requires at least two triggerable Brand Profile fields. A cell may therefore be MA-eligible but MS-ineligible, or MS-eligible but MA-ineligible.

D.5 Aspect-level disaggregation

For diagnostic purposes, sentiment may be disaggregated by brand attribute category: pricing sentiment, feature sentiment, support/service sentiment, positioning sentiment. Aspect-level disaggregation follows the same brand-directed definition. It is not required for headline MS reporting but is included in full diagnostic output where the data support it.

D.6 Reliability reporting

Inter-rater reliability for MS is reported with Krippendorff's alpha for ordinal data (positive > neutral > negative). The target is $\alpha \geq 0.75$ for the LLM evaluator against the human gold standard, consistent with §7.2. Because MS is interpretive, human-human agreement before adjudication is lower for MS than for MA. Both are reported and distinguished.

D.7 Positivity-offset test

Before production use, the evaluator is tested for systematic positivity lean on brand-directed sentiment specifically. The test compares the distribution of LLM evaluator classifications against human adjudicated classifications on the gold standard sample. A significant positive offset, assessed using a chi-square test, $p < 0.05$, requires correction before scores are published. Where the positivity-offset test is significant, correction is applied by recalibrating the evaluator against the human-adjudicated gold standard. The per-class confusion matrix from the gold-standard sample defines a calibration adjustment applied to subsequent classifications, and raw and calibrated MS are reported side by side for the affected period. The evaluator version in which the correction was applied is recorded.

Appendix E: Evaluator Prompt Templates

E.1 MA evaluator prompt

The MA evaluator is called once per eligible AI response to be scored. Cell-level MA values are then constructed from the response-level scores according to the rule in §5.3. The session must be fresh (no prior exposure to other responses in the same session) and the evaluator model must be from a different model family than the engine that produced the response.

Template:

```
You are an evaluator measuring how accurately an AI-generated brand description matches a declared
↪ Brand Profile.

Brand Profile:
[INSERT BRAND PROFILE AS STRUCTURED LIST: field name, declared value, field type]

AI response to evaluate:
[INSERT RESPONSE TEXT]

Instructions:
1. Identify every Brand Profile field that is directly addressed in the response. A field is
↪ addressed if the response makes a claim that can be compared to the declared value.
2. For each addressed field, assign a score using the rubric below. Do not score fields not
↪ addressed.
3. A response that addresses fewer than two fields is not eligible; output {"eligible": false,
↪ "reason": "..."}.
4. Output JSON only. Do not include prose outside the JSON.

Output format:
{
  "eligible": true,
  "triggered_fields": ["field_name_1", "field_name_2", ...],
  "scores": {"field_name_1": score, "field_name_2": score, ...},
  "alignment_score": mean_of_triggered_scores,
  "notes": "brief note on any ambiguous scoring decisions"
}

Scoring rubric:
[INSERT FULL RUBRIC FROM APPENDIX C §C.2]
```

E.2 MS evaluator prompt

The MS evaluator is called once per eligible AI response. Cell-level MS classifications are then constructed from the response-level classifications according to the rule in §5.3 and §6.3. It uses the same session-independence requirements as the MA evaluator.

Template:

```
You are an evaluator measuring brand-directed sentiment in an AI-generated response. You are
↪ classifying how the response characterises the brand, not the overall tone of the text.

Brand: [BRAND NAME]

IMPORTANT: Classify the sentiment expressed toward [BRAND NAME] specifically.
- A response that criticises the product category but praises [BRAND NAME] is POSITIVE.
- A response that is generally favourable but identifies [BRAND NAME] as a poor fit for certain
↪ users is NEGATIVE.
- Do not classify the overall text tone. Classify only the brand-directed characterisation.

AI response to evaluate:
[INSERT RESPONSE TEXT]
```

```
Is there any substantive evaluative characterisation of [BRAND NAME] in this response (beyond
 ↪  bare-name inclusion in a list)?
- If no: output {"eligible": false, "reason": "bare-name mention only"}
- If yes: classify as positive, neutral, or negative per the definitions below and output JSON.

Definitions:
- positive: approval, endorsement, recommendation, or favourable comparison toward [BRAND NAME]
- neutral: mention without valenced characterisation, or balanced characterisation with no net
 ↪  direction
- negative: disapproval, criticism, unfavourable comparison, or warning about [BRAND NAME]

Output format (when eligible):
{
  "eligible": true,
  "sentiment": "positive" | "neutral" | "negative",
  "dominant_characterisation": "one-sentence summary of what drives this classification",
  "notes": "note any ambiguity or mixed signals"
}
```

E.3 Ensemble protocol

Both MA and MS evaluations run as an ensemble of three independent evaluations per response (same prompt, same evaluator model, separate sessions). For MA, the headline alignment score is the median of the three scores. For MS, the headline classification is the majority vote across the three evaluations. Where three-way disagreement occurs (one positive, one neutral, one negative), the response is flagged for human review.

The evaluator model version and the evaluation date are recorded in the measurement log alongside each response score.

Appendix F: Diagnostic Outputs

The metrics tell a brand where it stands, and the diagnostics tell it what to do. The following standard diagnostic views are computed from data the measurement already collects, without additional machinery:

Prompt gap list: the prompts where the brand is absent or far below the category norm, grouped by query type (categorical, comparative, use-case, problem-solution). This names the questions the brand is losing in the category answer space.

Per-engine comparison: the per-engine score table read diagnostically. Engine underperformance identifies where to investigate after an engine-change discontinuity.

Field-level misalignment report: MA disaggregated by Brand Profile field. This converts a headline MA score into a work order: "pricing is the dominant misalignment; fix the cited sources stating the old tier."

Sentiment drivers: for negative and mixed bands, representative response excerpts per the codebook's categories, showing what AI says when it characterises the brand unfavourably.

Version-change decomposition: at a prompt-set version change, how much of the score movement is performance on familiar questions versus position in the new question space.

Working paper, June 2026. Methodological specification.