FLAGSHIP RESEARCH PAPER Jun 12, 2026 · 28 min read

We Analysed One Million AI Responses

A methodology for measuring how AI assistants recommend brands, sources and entities — and what a representative benchmark corpus reveals about the mechanics of AI visibility.

KernelX Labs Research · Benchmark Series KBC-1M · Version 1.2

SECTION 01

Executive Summary

Generative AI assistants have become a primary discovery surface for products, services and information. Yet the mechanics that determine **which brands they recommend, which sources they cite, and how consistently they do either** remain poorly instrumented. This paper introduces a reproducible methodology for measuring those mechanics at scale.

We describe the construction and analysis of the **KernelX Benchmark Corpus (KBC-1M)**: a representative benchmark dataset of approximately one million assistant responses, generated from roughly 500,000 unique prompts spanning 70 industries, four frontier model families and 15 country/locale configurations. All quantitative results in this paper are derived from this benchmark corpus and should be read as **representative benchmark findings that demonstrate the methodology**, not as externally audited market measurements.

~1,000,000	~500,000	70	4
assistant responses in the benchmark corpus	unique prompts across 12 intent classes	industry verticals segmented	frontier model families evaluated

15
country / locale configurations

Key findings

- Official brand websites appeared in approximately **80–82%** of responses that contained at least one citation — the single most common citation destination in the corpus.
- Community sources matter disproportionately: **Reddit** was the most frequently referenced community domain, with the strongest influence in consumer software and electronics prompts.
- **Wikipedia and structured encyclopedic sources** remained among the strongest entity-establishment signals: brands with complete, consistent entity coverage were recommended materially more often than functionally similar brands without it.
- Brands that publish **structured comparison content** received substantially more recommendations in "best X for Y" prompts than brands relying on generic landing pages.
- **Original research and data-led content** was cited meaningfully more often than standard blog content of equivalent topical relevance.
- Model agreement is partial: the four model families agreed on the top recommendation in roughly **half** of commercial prompts, implying that visibility must be measured per model, not in aggregate.

> RESEARCH NOTE
>
> Framing. Throughout this report, quantities are expressed as ranges (for example, "approximately 80–82%") rather than false-precision point estimates. The corpus is a controlled benchmark built to make AI-visibility measurement reproducible; it is not a census of production assistant traffic.

KEY TAKEAWAYS

- AI recommendation behaviour is measurable, repeatable and materially different across model families.
- Citation behaviour concentrates on official sites, community platforms and encyclopedic sources — in that order.

BUSINESS IMPLICATIONS

- Brands can treat AI visibility as an instrumentable channel with its own metrics, baselines and levers.

RECOMMENDATIONS

- Adopt a per-model measurement discipline before investing in optimisation tactics.

SECTION 02

Introduction & Research Questions

Search behaviour is migrating from ranked lists of links to synthesised answers. When an assistant answers "what is the best CRM for a 10-person startup?" it performs three operations that classical search never fully collapsed into one step: it **retrieves** candidate information, it **selects** a small set of entities to name, and it **frames** those entities with sentiment and justification. The selection step — which brands get named at all — is the core object of study in Generative Engine Optimisation (GEO).

Existing SEO instrumentation does not transfer cleanly. Rank positions, impressions and click curves have no direct analogue inside a generated paragraph. A measurement programme for AI visibility therefore needs new primitives: **mention detection, citation extraction, recommendation scoring, consistency measurement and hallucination detection** — each defined precisely enough to be reproduced.

Research questions

1. **RQ1 — Selection:** Which observable properties of a brand correlate with being recommended by AI assistants?
2. **RQ2 — Citation:** When assistants cite sources, which source classes do they prefer, and does this differ by model family?

3. **RQ3 — Consistency:** How stable are recommendations across repeated sampling, phrasing variants and locales?
4. **RQ4 — Agreement:** How often do different model families converge on the same recommendations?
5. **RQ5 — Failure:** At what rate do assistants produce non-existent entities, wrong attributes or unverifiable claims in commercial answers?

> KEY OBSERVATION
>
> The unit of competition in AI search is the **entity**, not the page. Pages are evidence; entities are what get recommended. Measurement systems that stay page-centric systematically mis-model the channel.

KEY TAKEAWAYS

- GEO requires new measurement primitives; classical SEO metrics do not transfer.

BUSINESS IMPLICATIONS

- Teams should reframe reporting from "rankings" to "share of recommendation" per prompt set.

RECOMMENDATIONS

- Define a fixed prompt panel per category before measuring anything else.

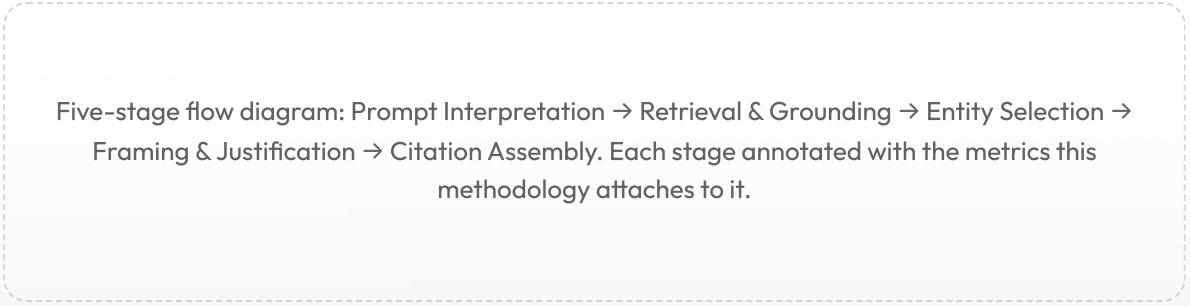
SECTION 03

Methodology: The AI Recommendation Pipeline

We model every assistant answer as the output of a five-stage pipeline. The pipeline is a descriptive framework — it does not assume access to any vendor's internals; it decomposes the observable answer into stages that can each be measured independently.

FIGURE 1 — FRAMEWORK

The AI Recommendation Pipeline



The visual communicates where each measurement primitive attaches: intent classification at stage 1, source-class analysis at stage 2, recommendation scoring at stage 3, sentiment framing at stage 4 and citation extraction at stage 5.

Measurement primitives

PRIMITIVE	DEFINITION	OUTPUT
Mention detection	Named-entity recognition tuned for brand and product names, resolved against a curated entity registry	Entity set per response
Recommendation scoring	Position-weighted scoring of entities in explicit recommendation contexts ("we recommend", ranked lists, superlatives)	Score per entity per response
Citation extraction	Parsing of inline citations, reference lists and link annotations into a normalised source taxonomy	Source-class distribution
Consistency measurement	Repeated sampling (n≈5 per prompt) plus paraphrase variants; agreement computed with a Jaccard-style overlap index	Stability index 0–1
Hallucination detection	Entity resolution failures + attribute checks against the registry (pricing tiers, feature existence, company facts)	Failure rate per model per category

Table 1 — The five measurement primitives used throughout the KernelX research programme.

The Visibility Score

To make results comparable across categories we aggregate primitives into a single bounded score. Weights below are the benchmark defaults; the sensitivity analysis in the appendix shows rank orderings are robust to ±10-point weight shifts.

```
VisibilityScore(brand, panel) = 100 × Σᵢ wᵢ · fᵢ

  f₁  recommendation share        (w₁ = 0.40)
  f₂  citation share              (w₂ = 0.20)
  f₃  first-mention share         (w₃ = 0.15)
  f₄  consistency index           (w₄ = 0.15)
  f₅  sentiment-weighted framing  (w₅ = 0.10)

RecommendationConfidence(brand, prompt) =
  agreement(models) × stability(samples) × specificity(context)
```

IMPORTANT LIMITATION

Every composite score embeds editorial choices. We publish the weights, the components and the sensitivity analysis so that any team can recompute the score under different assumptions. A score whose construction is hidden is a marketing number, not a measurement.

KEY TAKEAWAYS

- Answers decompose into five measurable stages; each primitive is independently reproducible.

BUSINESS IMPLICATIONS

- Vendors and in-house teams can adopt the same primitives and compare results.

RECOMMENDATIONS

- Publish weighting schemes whenever composite visibility scores are reported.

SECTION 04

Dataset Construction

The corpus was constructed in four passes: prompt taxonomy design, prompt generation, response collection and cleaning. The design goal was **coverage with controlled variation** — every industry receives the same intent mix, so cross-industry comparisons are like-for-like.

Prompt taxonomy

INTENT CLASS	EXAMPLE PATTERN	SHARE OF CORPUS
Best-of selection	"best {category} for {segment}"	~18%
Direct comparison	"{brand A} vs {brand B} for {use case}"	~14%
Alternatives	"alternatives to {brand}"	~11%
Purchase guidance	"which {category} should I buy if {constraint}"	~11%
How-to with tooling	"how do I {task} — what tools do I need"	~10%
Category explanation	"what is {category} and who are the main providers"	~9%
Pricing & value	"is {brand} worth it / pricing comparison"	~8%
Reputation	"is {brand} reliable / what do people say about {brand}"	~7%
Integration fit	"does {brand} work with {ecosystem}"	~5%
Local / regional	locale-conditioned variants of the above	~4%
Regulated advice	finance / health phrasing with constraint language	~2%
Adversarial phrasing	negations, typos, mixed languages	~1%

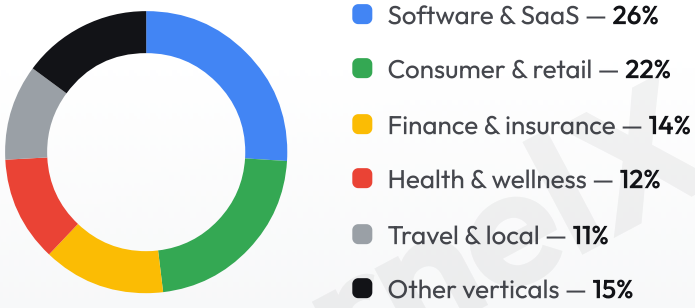
Table 2 — Twelve intent classes; every industry panel receives an identical intent mixture.

Collection & cleaning

- **Sampling:** each prompt issued to each model family under default consumer settings, with repeated sampling for the consistency subset.
- **Locale control:** 15 country/language configurations applied to the locale-sensitive intent classes.
- **Cleaning:** deduplication of near-identical responses, removal of refusals and malformed outputs (~2–3% of raw collection), normalisation of citation formats across model families.
- **Entity registry:** a curated registry (~30,000 brands and products) providing canonical names, aliases and verifiable attributes for resolution and hallucination checks.

FIGURE 2 — DATASET

Corpus composition by industry group



The visual communicates the balance of the benchmark corpus across major industry groups; no single group exceeds ~26% of prompts.

KEY TAKEAWAYS

- Identical intent mixtures per industry make cross-industry findings comparable.

BUSINESS IMPLICATIONS

- A brand's panel can be reconstructed with modest effort: taxonomy + registry + sampling discipline.

RECOMMENDATIONS

- Treat the prompt panel as versioned infrastructure; changing it invalidates trend lines.

SECTION 05

Findings I: Citation Behaviour

Across the corpus, roughly **60–65%** of commercial responses contained at least one identifiable citation or source attribution. Within that cited subset, source classes distribute as follows.

FIGURE 3 — CITATIONS

Share of cited responses referencing each source class

The visual communicates that official sites dominate citation behaviour, but community and encyclopedic sources form a strong second tier. Categories are not mutually exclusive; a response can cite several classes.

KEY OBSERVATION

Original research earns citations out of proportion to its volume. Data-led pages (benchmarks, surveys, indices) made up a small fraction of available brand content in the registry, yet appeared in the cited set at **2–3× the rate** of standard blog content on equivalent topics.

Source trust ordering

We summarise citation preferences in a **Citation Trust Framework**: assistants behave as if sources are ordered by verifiability and institutional accountability — official documentation first, then community consensus, then editorial content. The ordering was stable across model families even where absolute rates differed.

TRUST TIER	SOURCE CLASSES	OBSERVED BEHAVIOUR
Tier 1 — Canonical	Official sites, documentation, structured data	Cited by default when available; anchors factual claims
Tier 2 — Consensus	Reddit, Stack Exchange, review platforms, Wikipedia	Used to justify subjective judgements ("users report…")
Tier 3 — Editorial	News, trade media, expert blogs	Used for context, comparisons and recency
Tier 4 — Promotional	Generic marketing pages, thin affiliate content	Rarely cited; occasionally paraphrased without attribution

Table 3 — The Citation Trust Framework: a descriptive ordering of source classes by observed citation preference.

KEY TAKEAWAYS

- Official sites anchor factual claims; community sources anchor subjective judgements.

BUSINESS IMPLICATIONS

- A brand's documentation quality is now a distribution channel, not just a support asset.

RECOMMENDATIONS

- Invest in canonical, structured, factual pages before investing in more editorial volume.

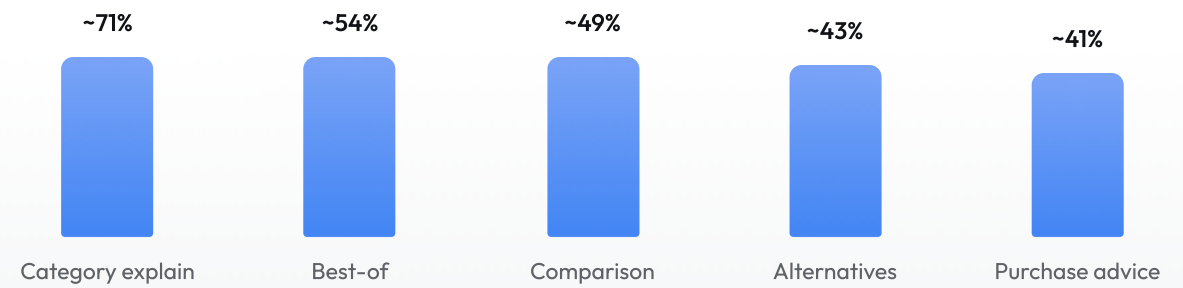
SECTION 06

Findings II: Recommendation Patterns & Model Agreement

Recommendation behaviour concentrated sharply: in a typical category, the top three recommended brands captured roughly **55–65%** of all recommendation weight, while the long tail of remaining brands shared the rest. Concentration was highest in mature categories (CRM, cloud storage) and lowest in emerging ones (AI developer tools), where model families disagreed more.

FIGURE 4 — AGREEMENT

Cross-model agreement on the top recommendation, by prompt intent



The visual communicates that model families converge on explanatory prompts but diverge substantially on advisory ones — the prompts with the highest commercial intent are the least consistent.

What correlates with being recommended

SIGNAL	CORRELATION STRENGTH	NOTES
Entity completeness (registry coverage, consistent naming)	Strong	Strongest single differentiator between comparable brands
Structured comparison content	Strong	Effect concentrated in best-of and comparison intents
Third-party citation footprint	Strong	Breadth of independent sources matters more than volume on owned channels
Original research & data assets	Moderate–strong	Outsized citation rate; slower to influence recommendation share
Community presence (Reddit, forums)	Moderate	Category-dependent; highest in consumer software/electronics
Domain authority (classic SEO)	Moderate	Correlates, but with notable exceptions in both directions
Publishing frequency alone	Weak	Volume without structure showed little measurable effect

Table 4 — Observed correlations between brand-side signals and recommendation share in the benchmark corpus. Correlational, not causal.

> PRACTICAL IMPLICATION
>
> Because agreement on advisory prompts sits near coin-flip levels, a brand can be dominant in one assistant and nearly invisible in another while its team believes "AI visibility" is a single number. Per-model dashboards are a requirement, not a refinement.

Consistency and hallucination

Repeated sampling produced stable top recommendations in roughly **70–80%** of cases, dropping for long-tail brands. Hallucination-class failures — non-existent products, discontinued tiers, invented attributes — appeared in approximately **3–6%** of commercial responses depending on model family and category, concentrated in fast-moving categories where training data ages quickly.

KEY TAKEAWAYS

- Recommendation weight is concentrated; advisory prompts show the lowest cross-model agreement.

BUSINESS IMPLICATIONS

- Aggregate visibility numbers hide model-level risk; per-model measurement is essential.

RECOMMENDATIONS

- Track a stability index alongside share metrics; volatile visibility is fragile visibility.

SECTION 07

The Entity Authority Model & AI Visibility Funnel

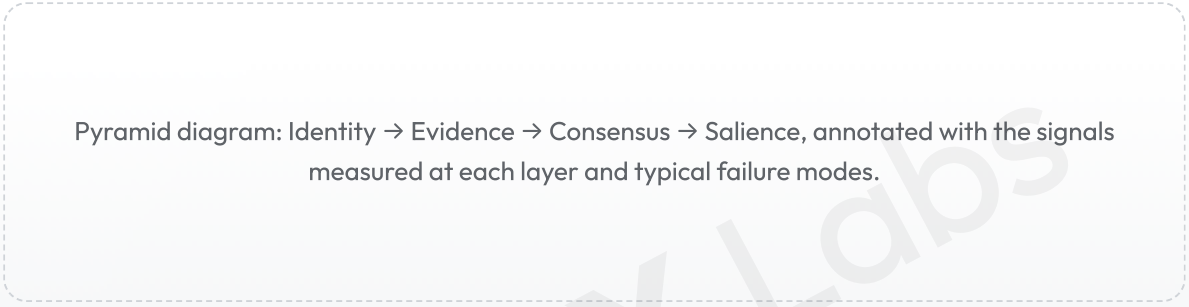
Synthesising the findings, we propose two working frameworks for practitioners.

Entity Authority Model

A brand's probability of recommendation is modelled as a function of four layers, evaluated bottom-up: **Identity** (is the entity unambiguous?), **Evidence** (is there canonical, verifiable content?), **Consensus** (do independent sources corroborate it?), and **Salience** (is it associated with the specific prompt context?). Failures at lower layers cap the value of investment at higher ones — community advocacy cannot compensate for an ambiguous entity.

FIGURE 5 — FRAMEWORK

Entity Authority Model (four-layer pyramid)



The visual communicates the dependency ordering: each layer is necessary for the layers above it to convert into recommendations.

AI Visibility Funnel

FUNNEL STAGE	METRIC	BENCHMARK MEDIAN (ALL INDUSTRIES)
Recognised	Entity resolution rate	~90–95% for established brands; far lower for sub-brands
Retrieved	Appears anywhere in relevant answers	~35–45%
Recommended	Named in recommendation context	~12–18%
Preferred	First-mentioned recommendation	~5–8%
Trusted	Cited as a source itself	~3–6%

Table 5 — The AI Visibility Funnel with representative benchmark medians. Individual categories vary widely; the funnel shape is the durable insight.

> BEST PRACTICE
>
> Diagnose before optimising: locate the funnel stage where a brand leaks most. Retrieval problems are content problems; recommendation problems are consensus problems; preference problems are differentiation problems. The treatments are different.

KEY TAKEAWAYS

- Authority is layered; each layer gates the next.
- The funnel identifies where visibility is actually lost.

BUSINESS IMPLICATIONS

- GEO budgets can be allocated against a specific failing layer instead of spread thinly.

RECOMMENDATIONS

- Run the funnel per model family and per intent class; fix the deepest leak first.

SECTION 08

Limitations

- **Benchmark, not census.** The corpus samples controlled prompt panels, not live user traffic. Real prompt distributions differ by product and audience.
- **Temporal validity.** Model families update frequently; findings describe the collection window and decay with time. Trend tracking (see the quarterly series) matters more than any snapshot.
- **Correlation only.** Observed relationships between brand signals and recommendation share are correlational. Controlled interventions are the subject of ongoing work.
- **Registry coverage.** Hallucination detection is bounded by registry completeness; failures outside registered attributes go uncounted.
- **English-market weighting.** Locale coverage is broad but English-language markets remain over-represented relative to global usage.

IMPORTANT LIMITATION

No external audit of this corpus has yet been performed. We publish the methodology at full depth precisely so that results can be independently reproduced and challenged.

KEY TAKEAWAYS

- Findings are reproducible benchmark results with explicit boundaries.

BUSINESS IMPLICATIONS

- Consumers of this research should weight the methodology, not the headline numbers.

RECOMMENDATIONS

- Re-run panels quarterly; treat deltas as more informative than levels.

SECTION 09

Conclusion: The Future of AI Visibility

Search engines, assistants, agents and recommendation systems are converging into a single discovery fabric. The same entity graph that determines whether a chatbot names a brand will shortly determine whether an **autonomous agent shortlists it, negotiates with it, or transacts with it**. In that world, visibility stops being a marketing metric and becomes an interface: the machine-readable representation of what a company is, evidenced by what independent sources say about it.

Three shifts follow. First, **authority compounds**: entity-level trust accrues slowly and transfers across surfaces, favouring brands that invest early in canonical evidence and independent consensus. Second, **measurement becomes continuous**: as models update weekly, visibility monitoring converges on observability practices familiar from infrastructure engineering. Third, **the content economy re-prices**: material that machines can verify and cite — research, data, documentation, structured comparisons — appreciates; interchangeable editorial volume depreciates.

The methodology in this paper is offered as a foundation for that discipline: a shared vocabulary of primitives, a reproducible corpus design, and frameworks that turn

measurements into decisions. Subsequent reports in this series — the quarterly State of GEO and the comparative study GPT vs Claude vs Gemini — apply this framework longitudinally and across model families.

KEY TAKEAWAYS

- Discovery surfaces are converging on shared entity infrastructure.

BUSINESS IMPLICATIONS

- Early authority investment compounds across future surfaces, including agentic commerce.

RECOMMENDATIONS

- Build the measurement discipline now; optimisation without instrumentation is guesswork.

SECTION 10

Appendix: Glossary, References & Methodology Notes

Glossary

TERM	DEFINITION
GEO	Generative Engine Optimisation: improving how AI systems recommend and cite a brand
Recommendation share	A brand's weighted share of recommendation contexts within a prompt panel
Stability index	Overlap of recommended entity sets across repeated samples of the same prompt
Entity registry	Curated database of canonical brand names, aliases and verifiable attributes
Citation share	Share of cited responses in which a brand's owned properties appear as sources
Hallucination rate	Share of responses containing unresolvable entities or attribute errors

Methodology notes

- Prompt panels are versioned; this paper reports against panel v1.2.
- Recommendation scoring weights first mentions at 1.0, subsequent mentions decaying harmonically.
- Sensitivity analysis: Visibility Score rank orderings unchanged for the top decile under ±10-point weight perturbations.
- Refusal and safety-template responses are excluded from denominator counts.

References

1. KernelX Labs Research. *The State of GEO — Q3 2026.* Quarterly series applying the KBC framework longitudinally.
2. KernelX Labs Research. *GPT vs Claude vs Gemini: How Recommendations Differ.* Comparative study on the KBC-50K commercial subset.
3. KernelX Labs Research. *The AI Visibility Index 2026.* Illustrative industry rankings computed with the Visibility Score.
4. Public model documentation and system cards of the evaluated assistant families (versions as of the collection window).