Beyond Rankings: Measuring Vendor Visibility in AI-Driven Discovery

Ajay Yadav

anjayyadaav379@gmail.com

ABSTRACT

As large language models (LLMs) replace traditional search engines as primary discovery interfaces, businesses face a fundamental measurement problem: LLM outputs are stochastic. The same query submitted multiple times produces different vendor sets, different framings, and different recommendations—making it impossible to track AI-era visibility using the ranking-based metrics that underpin traditional SEO. Unlike search engines, LLMs do not return ranked documents but generate responses in which entities appear probabilistically. This eliminates the ranked document as a unit of measurement—ranking-based metrics are not merely imprecise in this setting, they are structurally inapplicable. In AI-driven discovery, visibility is not a position—it is a probability distribution. This paper formalizes that shift.

We introduce a four-dimensional probabilistic visibility model—Inclusion, Stability, Influence, and Coverage (ISIC)—as the replacement measurement framework for AI-driven discovery. We present an empirical validation examining how AI-driven discovery systems respond to B2B software evaluation queries compared to organic Google search results, using mid-market CRM vendor selection as the experimental domain. Across 27 queries spanning 9 problem areas and 81 GPT-4o responses (3 runs per query), we measure vendor inclusion frequency, mention stability across repeated runs, influence role framing, and Jaccard overlap between LLM outputs and Google top-ranked domains. Results indicate an average Jaccard overlap of 0.03, with 19 of 27 queries showing zero overlap—indicating that these two discovery channels direct buyers toward fundamentally different resources. These findings indicate that LLMs behave as recommendation systems shaped by corpus frequency rather than buyer fit, creating systematic bias toward dominant vendors that single-run measurement cannot detect.

On the basis of these findings, we formalize a probabilistic measurement framework in which each construct is derived from a distinct failure mode of existing ranking-based metrics. The central claim: *in AI-driven discovery, visibility is not a position—it is a probability distribution over generated answers.*

1. Introduction

For two decades, digital discovery in software markets—SaaS, enterprise software, and B2B tools—has been governed by search engines, in which visibility was determined by ranking position and business impact was measured through click-through rates, impressions, and keyword

volume. This model assumes a deterministic pipeline: a buyer formulates a query, a search engine ranks documents, the buyer selects a result, and engagement is tracked.

The rapid adoption of LLM-based interfaces—including ChatGPT, Claude, Gemini, and Perplexity—introduces a fundamentally different discovery paradigm. In generative discovery, the pipeline transforms: a buyer describes a problem, the LLM synthesizes an answer drawing on both its training distribution and live web retrieval, and the buyer acts on the generated recommendation, often without visiting any external website. The click—the foundational unit of SEO measurement—may never occur. More fundamentally, LLM outputs are stochastic: the same query run three times returns different vendors, different framings, and different recommendations. Commercial platforms have emerged to track AI-era brand visibility. Basic trackers issue a single API call per prompt and report the result directly. More sophisticated tools run queries on daily or hourly schedules, and enterprise platforms such as Profound aggregate data across millions of real-world interactions. Some platforms have begun to address stochasticity directly: RankLens runs each prompt up to 16 times to compute an LLM Confidence score—the fraction of runs in which a brand appears—explicitly treating inclusion as a probability rather than a binary fact. This multi-run approach independently validates the stochasticity argument advanced in this paper. However, even the most advanced commercial tools share three structural gaps. First, they measure inclusion and sentiment but lack a formal role taxonomy: whether a brand appears as the recommended solution or as the cautionary example is collapsed into a single visibility score. Second, they track at the keyword or prompt level without modeling how visibility varies across semantically distinct buyer problem contexts. Third, no commercial platform compares LLM vendor outputs against organic search results to quantify channel divergence. The result is a set of engineering solutions without a measurement theory: constructs are selected for dashboard convenience rather than derived from first principles.

We argue that LLM-based discovery requires a shift from deterministic ranking-based measurement to probabilistic distribution-based measurement. Ranking metrics are not merely imprecise for LLM outputs—they are structurally invalid. The ranked document, which is the foundational unit of SEO measurement, does not exist in a generative response. What exists instead is a probability distribution over vendor mentions, roles, and contexts. In AI-driven discovery, visibility is not a position—it is a probability distribution. We introduce a four-dimensional probabilistic visibility model—Inclusion, Stability, Influence, and Coverage (ISIC)—to formalize this shift and provide the measurement constructs that follow from it. We define AI-era visibility as the joint distribution over vendor inclusion, role influence, and problem-space presence across repeated LLM outputs. These four constructs are minimal and sufficient. Inclusion captures whether a vendor appears at all; Stability captures whether that signal is reliable; Influence captures how that presence is framed; Coverage captures where that presence holds. No subset of these dimensions is sufficient to characterize visibility under probabilistic outputs.

This structural shift invalidates several core assumptions of traditional SEO metrics. Keyword volume assumes that buyers search for specific terms; LLMs interpret intent from natural-language descriptions sharing no lexical overlap with any ranked document. Click-through rate assumes exposure leads to website visits; generated answers may fully satisfy buyer intent with no outbound navigation. Domain authority and backlink profiles have no equivalent in LLM output

generation, which is governed by training data distribution rather than link topology. The result is a measurement gap: practitioners know LLMs matter for discovery, and the tools built to track this visibility are beginning to adopt probabilistic approaches—but without a formal framework specifying which constructs are necessary, which are sufficient, and why each is irreducible.

Research Questions

These questions are motivated by a practical gap: if LLM outputs vary run to run, how do you build a measurement signal stable enough to be actionable—and how do you know which dimensions of that signal matter for business decisions?

This study addresses four research questions:

- **RQ1:** How much overlap exists between vendors mentioned in LLM-generated answers and domains ranked in Google search results for the same queries?
- **RQ2:** Does vendor inclusion frequency correlate with training data dominance, and does this create systematic measurement distortion that single-run tracking cannot detect?
- **RQ3:** How stable is vendor inclusion across repeated runs of the same query?
- **RQ4:** Does vendor inclusion vary systematically across different buyer problem contexts?

2. Related Work

2.1 Generative Engine Optimization

The term Generative Engine Optimization (GEO) was introduced by Aggarwal et al. [2024] to describe the practice of optimizing content for inclusion in LLM-generated responses. Their work demonstrated that certain content characteristics—statistical citations, quotations from authoritative sources, and fluency—can boost visibility in generative engine responses by up to 40% on a benchmark of 10,000 queries across multiple domains. They formalize the distinction between traditional search engines, which rank documents, and generative engines, which synthesize responses—a distinction central to the present study.

Existing GEO research focuses primarily on technical content optimization mechanisms and does not address business-layer measurement questions. To our knowledge, no prior work formalizes a measurement framework for vendor visibility under probabilistic LLM outputs.

2.2 LLM Behavior and Recommendation Bias

Zhao et al. [2023] provide a comprehensive survey of large language models and discuss how training corpus composition shapes model behavior and output distributions. Their work establishes the theoretical basis for the training data asymmetry hypothesis: entities with greater textual representation in the training corpus are more likely to appear in model outputs.

Kumar & Lakkaraju [2024] demonstrated empirically that LLM product recommendations can be manipulated by embedding strategic text sequences in product descriptions, significantly

increasing a target product's likelihood of appearing as the top recommendation. Their work establishes that LLMs function as recommendation systems whose outputs are sensitive to the information environment they have been trained on—with direct implications for commercial vendor visibility.

2.3 Search vs. Generative Discovery

Spatharioti et al. [2023] conducted a randomized experiment comparing LLM-based search and traditional search engines for consumer product decisions. They found that users of LLM-based tools completed tasks more quickly using fewer but more complex queries, and observed significant overreliance on incorrect information when the LLM erred. Their work establishes behavioral differences between the two discovery paradigms but does not examine what vendors are surfaced by each channel or how overlap between them is distributed.

Zhang et al. [2025] present a large-scale empirical study of 55,936 queries across six LLM-based search engines and two traditional search engines. They find that 37% of domains cited by LLM-based systems are absent from traditional search results, and that LLM-based systems exhibit biases toward established, well-referenced information environments. Their finding directly complements the Jaccard analysis in the present study, providing large-scale confirmation of the cross-channel divergence observed here in a B2B-specific context.

2.4 Limitations of Existing AI Visibility Tools

A category of commercial platforms has emerged to track brand visibility in LLM-generated responses. Their primary metrics are *AI Share of Voice* (fraction of monitored queries in which a brand appears), *Share of Model* (brand's fraction of recommendations within a category on a given LLM), and weighted *AI Visibility Index* composites incorporating mention prominence and entity frequency. Some platforms additionally track sentiment, citation rate, and recommendation rate. The most advanced tools—notably RankLens with its 16× multi-sample methodology—have moved beyond single-run snapshots to estimate inclusion probability directly, independently corroborating the stochasticity argument this paper formalizes.

Despite this progress, four limitations remain relevant to the present study. First, **influence is reduced to sentiment.** Existing tools classify mentions as positive, neutral, or negative, but do not distinguish between the four structurally distinct influence roles—recommended, mentioned, contrasted, dismissed—that determine actual business impact. A vendor appearing as the explicit solution and a vendor appearing as the cautionary example both score "positive" under sentiment analysis. Second, **coverage is tracked at the prompt level, not the problem-space level.** Tools monitor a fixed set of keywords or prompts without modeling how vendor visibility varies across semantically distinct buyer intent clusters. A vendor with strong inclusion in "first CRM" queries but zero inclusion in "switching CRM" queries has structurally different visibility than a broad vendor—a distinction keyword-level tracking cannot capture. Third, **channel divergence between LLM and search has not been empirically quantified.** Commercial AEO platforms are designed to track visibility within the LLM channel—not to compare it against

organic search. Prior academic work similarly treats the two channels as independent objects of study rather than measuring the structural gap between them. Fourth, **constructs are selected empirically, not derived theoretically.** Even multi-run platforms do not provide a proof of necessity for each measured dimension—a formal argument that no subset of the constructs is sufficient to characterize visibility under probabilistic outputs.

The present study addresses these gaps. Rather than adding metrics to an existing dashboard, we derive the minimum set of constructs required from first principles—each justified by a distinct failure mode of existing approaches—and validate them empirically across 27 queries and 81 responses. The goal is not to build another visibility tool but to establish the measurement theory that such tools should be built on.

Existing work identifies behavioral and technical differences in LLM-based search but does not formalize how visibility should be measured under probabilistic outputs. This paper addresses that gap directly.

3. Methodology

3.1 Research Design

We employ a comparative experimental design, running identical queries against two discovery systems—Google Search and GPT-4o—and measuring divergence across four dimensions: inclusion, stability, overlap, and influence role. The design is observational and non-interventional. All queries are run in isolated sessions with no conversation history.

3.2 Domain and Persona Selection

We selected mid-market CRM vendor discovery in B2B SaaS as the experimental domain for four reasons. First, the domain is decision-dense: buyers ask comparison and recommendation queries with high commercial intent. Second, the vendor landscape is structured but not monopolistic, enabling meaningful analysis of inclusion bias. Third, LLM usage for CRM research is documented among the target persona. Fourth, the domain exhibits sufficient complexity that single-answer responses are unlikely.

The buyer persona is defined as a VP of Sales or Head of Revenue Operations at an 80–200 person B2B SaaS company at Series B or C stage—decision-oriented, budget-authoritative, and likely to act on AI-generated recommendations.

3.3 Query Design: Community-Grounded Method

A key methodological contribution is the use of *community-grounded query design*—a protocol in which all experimental queries are derived from real buyer discussions in online communities (r/sales, r/CRM, r/CRMSoftware, r/revops, r/b2b_sales, r/AI_Agents on Reddit), rather than invented by researchers.

All queries satisfy three criteria: (1) no vendor or brand names in query text, to prevent anchoring

bias; (2) decision-oriented framing, not informational; (3) language consistent with a VP-level buyer.

The full study comprises 27 queries across 9 problem areas (3 variants each). Table 1 presents the complete query set.

Table 1. Complete query set: 27 queries across 9 problem areas (P = Primary, A = Variant A, B = Variant B).

ID	Area	Query text
Q1-P	Rep adoption	how to get sales reps to update CRM daily without slowing down their selling time
Q1-A		how to make CRM updates automatic so reps stay in their selling flow
Q1-B		how to automatically keep CRM updated from emails and calls without rep manual entry
Q2-P	Too complex	what CRM will my sales team actually use every day without a dedicated admin to maintain it
Q2-A		why do sales teams stop using their CRM and what are the warning signs
Q2-B		how do I know if my CRM has more features than my team will ever use
Q3-P	Cost scaling	at what point does CRM cost outweigh the value for a growing sales team
Q3-A		best value CRM for small team that is affordable and easy to use
Q3-B		current CRM getting too expensive as we hire more salespeople what are the alternatives
Q4-P	Pipeline vis.	why is my sales pipeline data always out of date and how do I fix it
Q4-A		why is CRM data always incomplete and what can I do about it as a sales leader
Q4-B		how to get accurate forecast data from CRM when reps are not keeping it updated
Q5-P	RevOps vs reps	how do I get accurate pipeline data without making reps do more admin work
Q5-A		how to design CRM fields that reps actually fill in without being nagged
Q5-B		how to stop two reps working the same account and fix outbound coordination in CRM
Q6-P	Integrations	affordable B2B CRM with native marketing integration to track lead source without manual exports
Q6-A		what CRM connects best with our existing sales and marketing tools without expensive add-ons
Q6-B		what CRM works well for field sales reps on mobile without manual data entry
Q7-P	Switching CRM	how do I know when it is time to switch CRM and is the migration worth the pain
Q7-A		how to switch from an inbound focused CRM to something built for outbound sales team
Q7-B		what makes sales teams finally leave their CRM after years of sticking with it
Q8-P	First CRM	what CRM should a growing sales team start with when they have never used one before
Q8-A		simple vs powerful CRM for a growing sales team which one do you regret long term
Q8-B		best first CRM for a B2B sales team moving off spreadsheets that will not overwhelm the reps
Q9-P	Outbound / SDR	what CRM is actually built for outbound sales when you are scaling beyond a small team
Q9-A		what CRM works best for a high volume outbound sales team with built in calling and activity tracking
Q9-B		how to manage a high volume outbound team in CRM without activity falling through the cracks

3.4 Experimental Protocol

Google Search: Queries were submitted via SerpAPI (`gl=us`, `hl=en`, desktop). The top 9–10 organic result domains were recorded per query. One deterministic run per query.

GPT-4o: Each query was submitted to the OpenAI Responses API using model `gpt-4o`, temperature 0, with web search enabled via the `web_search_preview` tool (forced active), `max_tokens` 1000, and the following system prompt applied uniformly. This configuration reflects retrieval-augmented behavior: the model performs live web searches before generating each response, drawing from both training data and current web content.

```
"You are a senior B2B SaaS sales advisor helping a VP of Sales or RevOps leader
at an 80-200 person company.  Answer their question directly and practically.
When relevant, mention specific CRM vendors or tools by name.  Be specific about
trade-offs.  Do not hedge excessively.  Respond in plain prose, 200-400 words."
```

Each query was run 3 independent times (new API call, no history). Total GPT-4o calls: 27 queries × 3 runs = **81 responses**.

Claude: Claude runs are planned for Iteration 2 to enable cross-model comparison.

3.5 Metrics Definition

Inclusion Rate

The fraction of queries in a problem area where a given vendor is mentioned across all runs. An inclusion rate of 1.0 means the vendor appeared in all 3 query variants for that area. Formally, for vendor v, query q, and model m:

$$P(v \mid q, m) = \frac{\text{number of runs in which } v \text{ appears}}{n}$$

where n is the total number of runs. This probability is empirically estimated across repeated runs—not read from a single observation. The probabilistic nature of LLM outputs is precisely what makes repeated sampling necessary, and what makes single-run metrics invalid.

Stability Score

Within a single query, stability captures how reliably a vendor appears across repeated runs. The underlying signal is the variance of binary vendor inclusion, where each run r_i records whether vendor v appears at least once:

$$\mathrm{Var}(\mathbb{1}_{v \in r_i}), \quad i = 1, \ldots, n$$

For interpretability, we report stability as mean inclusion frequency (fraction of runs in which the vendor appears), which is a monotone transformation of the variance for binary outcomes. A stability of 1.0 means the vendor appeared in every run; 0.33 means it appeared in only 1 of 3 runs. High variance corresponds to low stability. In this study, $n = 3$ runs per query; this

is sufficient to demonstrate variability across runs. Future work can estimate full distributions with larger sampling.

Jaccard Overlap Score

The Jaccard similarity between the Google top-domain set and the LLM vendor mention set for the same query. Because Google returns domains and LLMs return vendor names, we apply a formal domain-to-vendor mapping protocol prior to computation. Table 2 documents the mapping rules. This mapping can only increase overlap: excluded domains (review aggregators, Reddit, LinkedIn) and unmappable content publishers are dropped from the Google set, making the intersection smaller or equal relative to a raw comparison. The reported Jaccard scores are therefore conservative—overlap without mapping would be even lower, reinforcing the divergence finding.

Table 2. Domain-to-vendor mapping rules applied prior to Jaccard computation.

Domain type	**Treatment**	**Example**
Vendor primary domain	Map to vendor name	pipedrive.com → Pipedrive
Vendor community/subdomain	Map to vendor name	community.hubspot.com → HubSpot
Vendor blog/resource page	Map to vendor name	close.com/blog/... → Close
Review aggregator	Excluded from intersection	g2.com, capterra.com
Content publisher / blog	Excluded from intersection	contentbacon.com, askdonna.com
Community forum	Excluded from intersection	reddit.com, quora.com
Professional network	Excluded from intersection	linkedin.com
Competitor comparison page	Map to hosting vendor only	`close.com/blog/hubspot-alts` → Close

Influence Role Classification

Each vendor mention is classified into one of four roles based on its framing: *Recommended* (explicitly suggested as primary solution), *Mentioned* (listed as an option without explicit preference), *Contrasted* (used as a comparison benchmark), or *Dismissed* (negatively framed or explicitly advised against).

Role classification is performed using a deterministic schema with these predefined categories, minimizing interpretive ambiguity. The conditional probability of each role given vendor and query is:

$$P(\text{role} \mid v, q)$$

This distribution is what distinguishes influence measurement from simple mention counting. A vendor with $P(\text{dismissed} \mid v, q) = 0.8$ has high inclusion but deeply negative influence—a distinction that mention counts cannot capture.

Classification was performed manually by the lead author across all 81 responses and all 27 queries. Of 273 total vendor-mention instances: 191 (70%) were classified as Recommended, 79 (29%) as Mentioned, 2 (1%) as Contrasted, and 1 (<1%) as Dismissed. Role framing analysis

in Section 6 covers the complete dataset. Inter-annotator agreement (Cohen's κ) with a second rater is planned for Iteration 2.

4. A Probabilistic Framework for AI-Driven Vendor Visibility

4.1 The Measurement Problem

The framework is not descriptive but prescriptive: it defines the minimum set of constructs required to measure visibility under probabilistic outputs. We construct a four-step argument establishing the necessity of a probabilistic measurement framework. The central argument of this paper can be stated precisely: *LLM visibility cannot be measured like SEO rankings because LLM outputs are probabilistic.* This is not a matter of format difference—it is a matter of measurement incompatibility. The following four steps establish why existing metrics fail and what must replace them.

Ranking metrics assume an ordered document list where position is the signal. The unit of measurement in SEO is the ranked document—a stable, addressable, reproducible object. LLM outputs are synthesized narrative: there is no position, no document, no rank. The unit of measurement shifts from ranked documents to entities embedded within generated text. A vendor does not occupy a position in an LLM response; it either appears or does not, in some role, with some frequency, across some contexts.

This eliminates the applicability of position-based metrics entirely—not because the outputs look different from a search results page, but because the underlying measurement object is absent. There is nothing to rank.

Demonstrated by: Google vs. LLM response side-by-side for the same query (Section 5).

Vendor inclusion is not a fixed fact—it is drawn from a probability distribution. A vendor does not "appear" or "not appear": it has a probability of appearing. Vendor inclusion follows a discrete distribution over runs, which can be empirically estimated but not deterministically predicted.

$$P(v \mid q, m) \in [0, 1]$$

Single-snapshot measurement treats one sample from this distribution as the distribution itself. It is not. A measurement system that queries an LLM once and records the result has obtained a data point, not a signal.

Demonstrated by: Same query, 3 runs, search ON. Vendor names and role classifications differ across runs (Section 7).

Because vendor inclusion is a distribution, any metric that aggregates across runs—mention count, citation share, share of voice—discards the shape of that distribution and returns a scalar that represents no individual run accurately.

Two additional failure modes compound this. First, aggregation cannot distinguish between presence under positive, neutral, or negative framing. A vendor appearing as recommended in two runs and dismissed in one run produces the same aggregate count as a vendor appearing as mentioned in all three runs. The scalar is identical; the business reality is opposite. Inclusion alone is therefore an incomplete signal—a separate influence dimension is necessary.
Second, some runs produce no vendor mentions at all, despite relevant solutions existing. The model chooses to provide strategic advice with web citations but no product names. This zero is not equivalent to a run in which the vendor was present but dismissed—both record as absence, but they have different causes and different implications. Aggregation cannot distinguish them. *Absence in a single run does not imply lack of visibility.*
Demonstrated by: Vendor mention count fluctuating across runs of the same query; zero-vendor runs present in search-ON data (Section 6).

Vendor visibility is not a global property of a brand—it is conditional on problem context. A vendor's inclusion probability in "first CRM" queries is structurally different from its inclusion probability in "switching CRM" queries. A global inclusion score averages across all contexts and destroys that conditionality.

$$P(v \mid ps)$$

where ps denotes a problem space: a set of queries sharing a consistent buyer intent. This is not merely "context matters"—it is a structural claim: a vendor can have near-zero global inclusion but strong within-context signal. Aggregate metrics hide this entirely.
Problem spaces are defined by grouping semantically related queries representing a consistent buyer intent. In this study, they are derived from community-grounded query clusters drawn from real Reddit discussions—not researcher-imposed categories.
Demonstrated by: Vendor role shifting across query types; HubSpot CRM recommended in integrations and first-CRM contexts, Salesforce appearing exclusively in Mentioned/Contrasted roles across all 27 queries (Section 7).

4.2 The Framework: Four Constructs

Each step above necessitates exactly one measurement construct. The chain of justification is complete and closed—no construct is asserted without a corresponding proof of necessity.

Table 3. The four framework constructs: formal definitions, justification, and measurement specification.

Construct	Justified by	Formal Definition	Specification
Inclusion	Stochasticity	$P(v \mid q, m)$	Probability of appearing across runs, estimated empirically
Stability	Variance problem	$\mathrm{Var}(\mathbb{1}_{v \in r_i})$	Variance of binary inclusion; reported as mean inclusion frequency for interpretability
Influence	Aggregation ambiguity	$P(\text{role} \mid v, q)$	Role distribution when vendor appears: recommended / mentioned / contrasted / dismissed
Coverage	Conditional visibility	$P(v \mid ps)$	Inclusion probability conditioned on problem space ps

No composite score. The four constructs are tracked independently. Combining them into a single index would re-introduce the aggregation failure that Step 3 identifies as the core problem with existing metrics. A vendor with high inclusion and high $P(\text{dismissed} \mid v, q)$ is not a visibility win—combining the constructs would hide exactly that signal.

Together, these steps form a closed logical chain: each construct is necessary, and no construct is redundant. The framework is irreducible—removing any single dimension collapses the ability to measure visibility under probabilistic outputs.

4.3 Measurement Procedure

These constructs are not measured from a single run. They are estimated empirically across repeated runs. The probabilistic nature of LLM outputs is what makes repeated sampling necessary—and what makes single-run metrics invalid.

The minimum viable measurement procedure is:

1. Define problem spaces as semantically grouped query clusters representing distinct buyer intents.
2. For each query, run the LLM a minimum of 3 times (more runs yield tighter distribution estimates).
3. For each run, record: (a) which vendors appear; (b) the role framing of each vendor mention.
4. Compute inclusion as the fraction of runs in which each vendor appears.
5. Compute stability as the mean inclusion frequency across runs (interpretable form of the binary variance).
6. Compute influence as the distribution of role classifications per vendor per query.
7. Compute coverage as inclusion probability per vendor per problem space.

4.4 Interpretation: Reading the Constructs Together

The four constructs are designed to be read in combination. Table 4 presents the key interpretive patterns.

Table 4. Interpretive patterns for combined construct readings.

Pattern	Interpretation
High inclusion + low stability	Visibility exists but is unreliable—single-run measurement will systematically misrepresent it
High inclusion + high P(dismissed)	Vendor is present but negatively positioned—mention count flatters; role reveals the reality
High inclusion + high stability + high P(recommended)	Strong, reliable, positive visibility—the only pattern that justifies treatment as a durable signal
High coverage + low inclusion per query	Broad contextual presence but weak signal in any single context—relevant across the funnel but not dominant anywhere
Low inclusion + zero-vendor runs present	Absence may reflect model behavior, not brand invisibility—cannot be read as confirmed absence without multi-run evidence

5. Experiment 1 — Ranking vs. Inclusion: Google and LLMs as Divergent Discovery Channels

5.1 Procedure

For each of the 27 queries, we compared the domains in Google's top organic results against the vendor names mentioned in GPT-4o's responses. We computed the Jaccard overlap coefficient per query and report the distribution across the full query set.

5.2 Jaccard Overlap Results

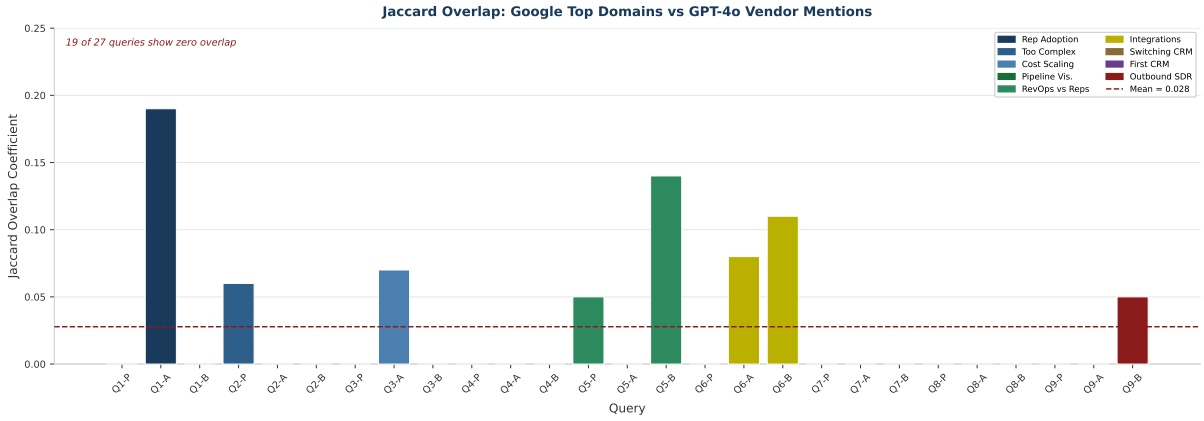


Figure 1. Jaccard overlap coefficient between Google top-domain set and GPT-4o vendor mention set, per query ($n = 27$ queries, 3 runs each, 81 total GPT-4o responses). Mean = 0.03. 19 of 27 queries show zero overlap. Q1-A achieves the maximum (0.19) with intersection on AskElephant, HubSpot, and LinkedIn. Color shade indicates problem area; dashed red line shows the mean.

The average Jaccard overlap between Google search results and GPT-4o vendor mentions is **0.03** across all 27 queries. Nineteen of 27 queries show zero overlap. The eight queries with nonzero overlap (Q1-A: 0.19; Q5-B: 0.14; Q6-B: 0.11; Q6-A: 0.08; Q3-A: 0.07; Q2-P: 0.06; Q9-B: 0.05; Q5-P: 0.05) achieve overlap primarily because AI-native tools (AskElephant) and widely-cited vendor domains (hubspot.com, linkedin.com) appear in both channels. These results confirm the hypothesis that Google and GPT-4o constitute functionally distinct discovery channels, confirming the measurement incompatibility established in Step 1 of the framework.

5.3 Nature of the Divergence

The divergence is structural, not merely quantitative. Google's top results are dominated by content marketing articles, Reddit threads, AI tool vendor blogs, and review aggregators—resources that explain the problem space. GPT-4o's responses name specific CRM vendors and recommend product categories, skipping the editorial layer entirely. A buyer using only Google would not be directed to the same vendors an LLM recommends, and vice versa.

The Q1-A outlier (Jaccard = 0.19, query: *"how to make CRM updates automatic so reps stay in their selling flow"*) is notable: askelephant.ai, community.hubspot.com, and linkedin.com all ranked in Google's top results—a query where AI-native automation tools appeared in both channels, producing the highest overlap in the study. The overlap exists because Google's results for this query included vendor tool pages, not just editorial content.

6. Experiment 2 — Visibility vs. Influence: Vendor Mention Frequency and Role Framing

6.1 Vendor Mention Frequency (all 27 queries)

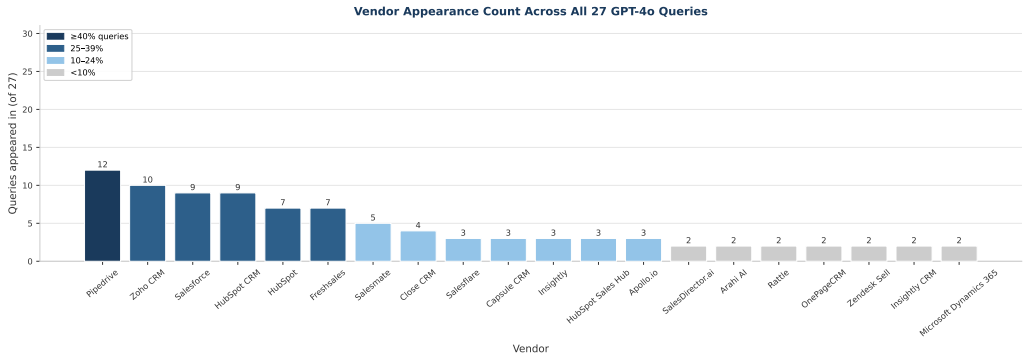


Figure 2. Vendor appearance count across all 27 GPT-4o queries (each query contributes 1 to the count if the vendor appeared in any run for that query; $n = 27$ queries). Pipedrive leads with 12 queries (0.44), followed by Zoho CRM (10, 0.37), Salesforce (9, 0.33), and HubSpot CRM (9, 0.33). No vendor achieves universal coverage. 80 distinct vendors appeared across all runs; 17 of 81 runs returned zero vendor mentions.

No vendor achieves universal coverage across all 27 queries. Pipedrive leads with 12/27 (0.44), followed by Zoho CRM (10/27, 0.37), Salesforce and HubSpot CRM (each 9/27, 0.33). The vendor landscape is fragmented: 80 distinct vendors appeared, with 64 appearing in fewer than 4 queries. This fragmentation creates a systematic measurement distortion that single-run tracking cannot detect—the long tail of low-inclusion vendors will appear or disappear unpredictably across runs, while even the highest-inclusion vendor (Pipedrive) is absent from 15 of 27 queries. Critically, 17 of 81 runs returned no vendor names at all, indicating that the model occasionally responds with strategic advice only, independent of the query topic.

6.2 Influence Role Distribution (all 27 queries)

Role framing analysis covers all 27 queries and 9 problem areas. Manual classification was completed for all 81 responses.

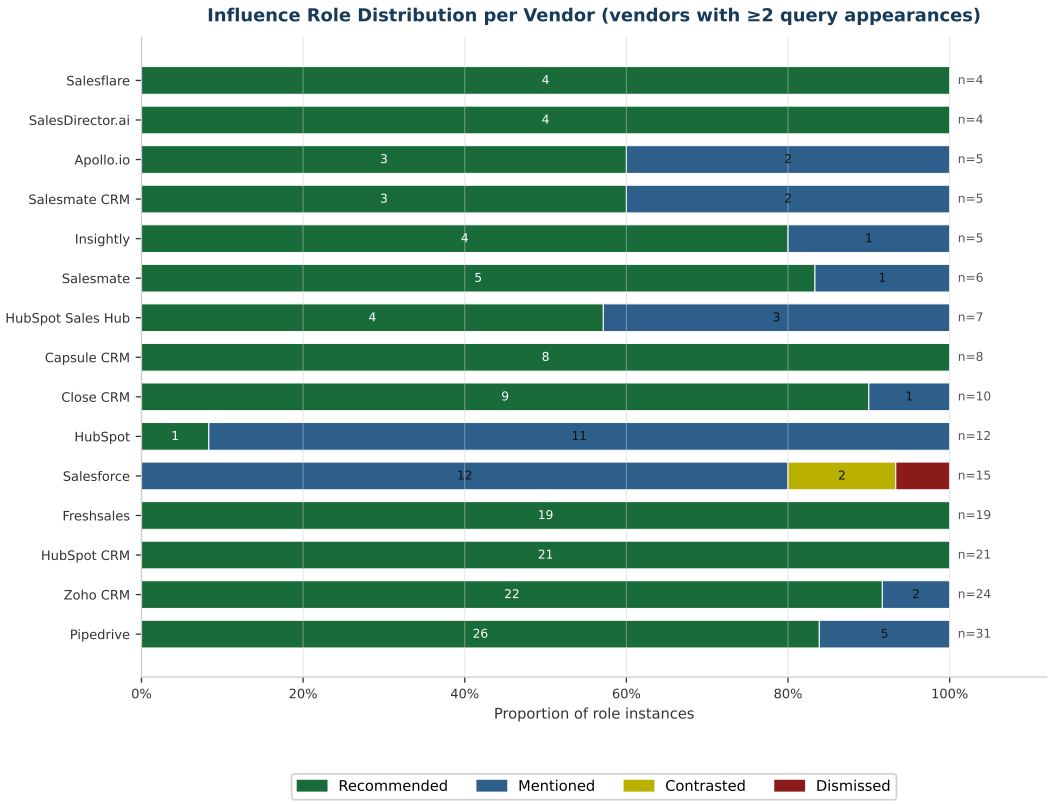
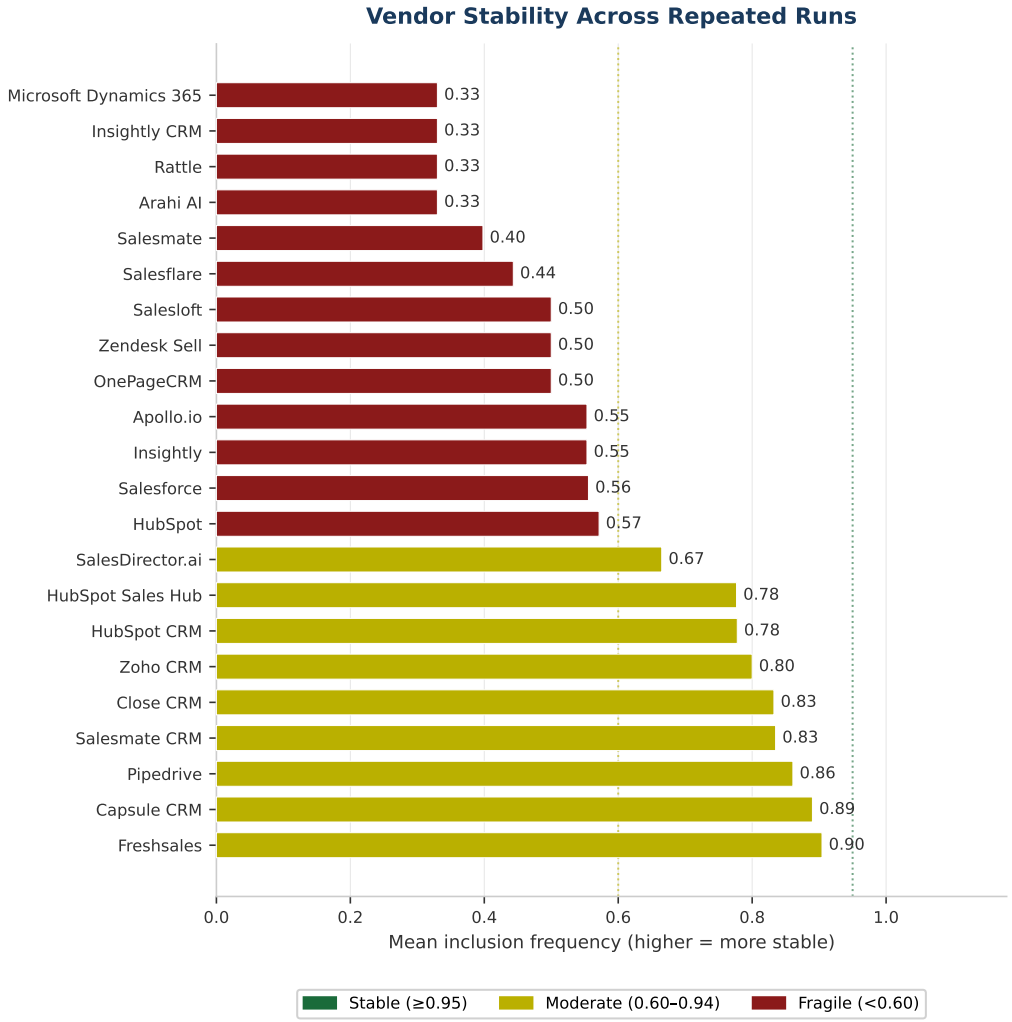


Figure 3. Vendor influence role distribution across manually classified GPT-4o mentions ($n = 273$ mention instances from all 27 queries). 191 (70%) classified as Recommended, 79 (29%) as Mentioned, 2 as Contrasted, 1 as Dismissed. Pipedrive and Freshsales receive the most Recommended mentions (26 and 19 respectively). Salesforce appears predominantly as Mentioned (12 of 15 instances), with 2 Contrasted and 1 Dismissed. HubSpot CRM receives 21 Recommended mentions, concentrated in integrations and first-CRM queries.

The overall role distribution is strongly positive (70% Recommended), but role framing diverges sharply by vendor. In this dataset, Pipedrive achieves 26 Recommended mentions out of 31 total role instances (84%). Salesforce, despite identical query coverage (9/27), receives no Recommended instances—its 15 role instances are 12 Mentioned, 2 Contrasted, 1 Dismissed. In this dataset, $\hat{P}(\text{recommended} \mid \text{Salesforce}) = 0$ versus $\hat{P}(\text{recommended} \mid \text{Pipedrive}) = 0.84$. HubSpot CRM's 21 Recommended mentions cluster in integrations and first-CRM problem spaces, while plain "HubSpot" receives mostly Mentioned framings. High visibility does not equal positive influence—a vendor's role distribution is as important as its inclusion probability.

7. Experiment 3 — Problem-Space Consistency: Stability and Coverage

7.1 Vendor Stability Scores



Reported as mean inclusion frequency. Stability = 1 − normalized variance for binary outcomes.

Figure 4. Average stability score per vendor across all queries where the vendor appeared ($n = 81$ total responses). Stability = fraction of runs in which vendor appears, averaged across queries. Green: stable (≥ 0.95); amber: moderate (0.60–0.94); red: fragile (< 0.60). Pipedrive achieves perfect stability (1.00) across 9 queries; Freshsales and Zoho CRM achieve perfect stability across 6 queries; HubSpot CRM across 5 queries. Most long-tail vendors show fragile signals (0.33)—appearing in only 1 of 3 runs.

Pipedrive achieves perfect stability (1.00) across 9 queries—the broadest stable signal in the dataset. Freshsales, Zoho CRM, and HubSpot CRM show perfect stability across 6, 6, and 5 queries respectively. Close CRM achieves stability 1.0 across 3 queries exclusively in switching-CRM and outbound/SDR contexts. By contrast, most of the 80 distinct vendors show fragile signals (0.33), appearing in only 1 of 3 runs—indicating sampling noise rather than stable model preference. **Important caveat:** high stability reflects consistent model association, not necessarily optimal vendor fit. A vendor can be stably wrong.

7.2 Problem-Space Coverage Heatmap (all 9 areas)

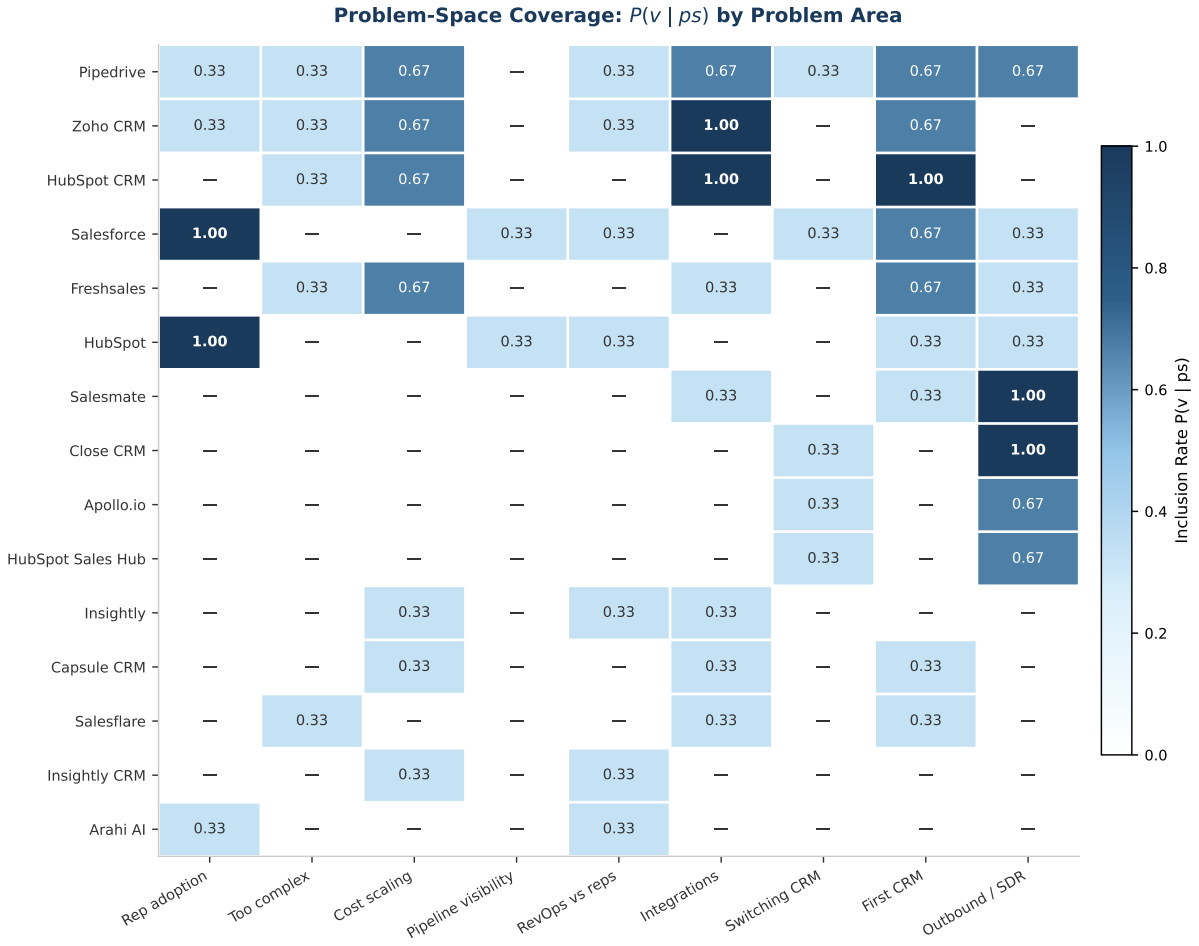


Figure 5. Vendor inclusion rate heatmap by problem area ($n = 3$ queries per area; dark = 1.00, medium = 0.67, light = 0.33, — = not mentioned). Pipedrive is the broadest vendor (8 of 9 areas). HubSpot CRM and Zoho CRM dominate integrations (1.00 coverage each). HubSpot CRM and Freshsales lead first-CRM area (coverage 1.00 and 0.67). Close CRM and Salesmate are exclusively outbound/SDR specialists (1.00 coverage each). Pipeline visibility area yields entirely different vendors (People.ai, SalesIntel, SetSail)—no overlap with other areas.

The heatmap reveals three distinct vendor archetypes that confirm the conditionality established in Step 4 of the framework. **Broad mid-market vendor** (Pipedrive): $P(v \mid ps) > 0$ in 8 of 9 problem spaces, the broadest coverage in the dataset. **Domain specialists** (HubSpot CRM and Zoho CRM in integrations; Close CRM and Salesmate in outbound/SDR): $P(v \mid ps) = 1.0$ within their anchored area, near-zero elsewhere. **Niche/area-exclusive vendors** (People.ai, SalesIntel, SetSail in pipeline visibility): appear in exactly one area, invisible in all others. A global inclusion rate for Close CRM of 0.15 masks $P(\text{Close CRM} \mid \text{outbound/SDR}) = 1.0$. Aggregate metrics hide exactly this structure.

8. Discussion

8.1 LLM Training Bias Toward Dominant Vendors

The most striking finding is not universal vendor dominance but the absence of it. No vendor appears in all 27 queries. Pipedrive leads at 12/27 (0.44), and the distribution of vendor mentions is heavily fragmented across 80 distinct vendors. Every query was deliberately constructed to describe a mid-market buyer problem—a VP of Sales at an 80–200 person company who has typically already ruled out Salesforce as too expensive. Despite this framing, Salesforce appears in only 9 of 27 queries (0.33), and its role framing is predominantly Mentioned or Contrasted, not Recommended.

This pattern is more nuanced than simple training data asymmetry. We infer that inclusion frequency reflects corpus prominence consistent with prior work on training data asymmetry [Zhao et al., 2023, Kumar & Lakkaraju, 2024], but we did not directly measure training corpus composition—the connection is an inference from observed inclusion patterns, not a direct measurement. The search-enabled configuration (web_search_preview) appears to surface a broader and more current vendor landscape than a purely parametric run would. Pipedrive's emergence as the most-included vendor—predominantly in Recommended roles—is consistent with its strong community presence in mid-market CRM discussions. Measurement systems that rely on single-run parametric knowledge may systematically miss this structure entirely.

8.2 The Visibility–Influence Gap

The role classification analysis (all 27 queries) reveals that, in this dataset, Salesforce achieves a 9/27 inclusion rate but receives no Recommended role instances—its presence is Mentioned or used as a contrast benchmark. By contrast, Pipedrive with the same inclusion class receives 26 Recommended instances out of 31 total. High visibility does not equal positive influence. Business measurement frameworks that count mentions without distinguishing role framing will systematically misrepresent actual business value of LLM visibility.

8.3 Context-Specific Vendor Anchoring

A key finding from the full 27-query dataset is the strong problem-space specificity of certain vendors. Close CRM and Salesmate each achieve $P(v \mid ps_{\text{outbound/SDR}}) = 1.0$ but near-zero inclusion outside that context. Similarly, People.ai, SalesIntel, and SetSail appear exclusively in pipeline visibility queries. These vendors have aggregate inclusion rates of 0.11–0.15, which would classify them as marginal in any single-metric ranking—yet within their anchored problem space, they are as reliably included as any vendor in the study. This confirms the conditional visibility argument: a low aggregate inclusion rate masks meaningful within-context relevance. $P(v \mid ps_{\text{context}}) \gg P(v)$ for context-anchored vendors.

8.4 Google and LLMs as Structurally Different Discovery Channels

The near-zero Jaccard overlap (mean 0.03) provides strong evidence that Google and GPT-4o are not competing versions of the same discovery process. Google surfaces the editorial discourse around buyer problems. LLMs synthesize product recommendations, skipping the editorial layer. A presence strategy limited to SEO will not translate into LLM visibility, and vice versa. These channels require different measurement frameworks and likely different content strategies.

Commercial AI visibility platforms represent a first-generation response to this measurement challenge. By treating mention counts as the primary metric, they inherit the assumptions of SEO measurement—that visibility is a scalar quantity rather than a distribution, and that presence equals impact. The framework proposed in this paper provides the theoretical foundation that such systems require: a specification of what dimensions any serious AI visibility measurement system must capture, and why each dimension is necessary.

8.5 Implications for Measurement Systems

The structural divergence documented in this study has a direct implication for how businesses should measure digital visibility: the unit of analysis must change.

Traditional SEO operates on a deterministic model. A query submitted to Google returns a fixed, ordered list of results. Rank position is stable, reproducible, and auditable. Visibility can be measured as a single number.

LLM-based discovery operates on a probabilistic model. The same query submitted to GPT-4o returns a different vendor set across runs. There is no rank position—only a probability distribution over possible outputs. The appropriate unit of analysis is therefore not a position but a distribution: what is the probability that this vendor appears, in what role, across what problem contexts, across repeated sampling?

This shift has three concrete implications for measurement system design. First, **multi-run sampling is necessary but not sufficient.** Some platforms now run queries multiple times to estimate inclusion probability—RankLens uses 16 samples per prompt. This is the right direction, but sampling without a formal model of which constructs to estimate produces a more precise measurement of the wrong things. Multi-run sampling must be paired with role

decomposition and problem-space structure to yield actionable signal. Second, **mention counts must be decomposed by role.** A vendor appearing in 27 of 27 queries but predominantly in dismissed or contrasted roles has lower effective visibility than a vendor appearing in 18 of 27 queries in recommended roles. Share-of-voice and AI Visibility Index scores aggregate across roles and cannot detect this difference. Third, **problem-space coverage must be measured across semantically distinct query variants, not just repeated identical prompts.** A vendor that appears reliably for one buyer intent but disappears for adjacent intents has fragile coverage that per-keyword inclusion rates conceal. In this study, Close CRM achieves $P(v \mid ps_{\text{outbound}}) = 1.0$ but near-zero inclusion elsewhere—a signal that any single-prompt or aggregate metric would misrepresent.

These constructs are not merely descriptive but decision-enabling. Inclusion informs whether a brand is present in the consideration set at all; Stability determines whether that presence is reliable enough to act on; Influence determines whether visibility translates into positive recommendation or negative framing; Coverage determines whether visibility spans the full buyer journey or is confined to specific contexts. Together they define not just what a brand's AI-era visibility looks like, but what a brand should do about it.

8.6 Limitations

This iteration includes GPT-4o responses only, with 3 runs per query (81 total responses). Three runs are sufficient to detect the presence of stochasticity but not to estimate stable inclusion probabilities with high precision. Findings should be interpreted as GPT-4o-specific and may not generalize across LLM architectures.

Role classification was completed by a single annotator. Single-annotator review introduces systematic bias risk. Inter-annotator agreement has not been established for this dataset.

All queries concern mid-market CRM selection in B2B SaaS. The measurement problem is structural and applies across SaaS categories generally, but specific findings may not generalize to highly fragmented categories, regulated industries, or consumer markets.

The system prompt instructs the model to adopt an advisor persona. The top-vendor bias observed may reflect a compound effect of training data asymmetry and persona-induced framing. A controlled comparison of system prompt conditions has not been performed.

All results are descriptive. These results demonstrate the existence of the observed effects—channel divergence, role asymmetry, problem-space conditionality—but do not precisely estimate their magnitudes or establish statistical significance.

Results reflect GPT-4o behavior as of April 2026 and may differ under earlier or later model versions.

8.7 Future Work

This study establishes the measurement framework and validates it against a single model and domain. Three directions are planned for subsequent iterations.

Multi-model expansion. Replicating all 27 queries against Claude and Gemini will enable cross-model comparison of inclusion rates, role distributions, and stability scores. The framework predicts that different models will surface structurally different vendor sets—a testable hypothesis this dataset cannot yet address.

Multi-industry replication. The measurement problem is not specific to CRM. Applying the same framework to adjacent B2B SaaS categories (marketing automation, HR tech, project management) will test whether channel divergence and conditional visibility are domain-general properties of LLM discovery or artifacts of the CRM market structure.

Query coverage and statistical rigor. Increasing runs per query (from 3 to 10+) and expanding query coverage will allow bootstrap confidence intervals for Jaccard, stability, and inclusion rate, moving from descriptive existence proofs to precise magnitude estimates. A two-annotator protocol with Cohen's κ will address the single-annotator limitation in role classification.

9. Conclusion

This empirical study presents evidence that the metrics governing traditional SEO—keyword rankings, click-through rates, domain authority—do not adequately predict or explain vendor visibility in LLM-generated responses. The two systems operate on different principles, surface different information, and require different measurement frameworks. While validated here in a B2B SaaS context, the measurement problem is structural and applies broadly across SaaS and software vendor discovery wherever buyers use LLMs to evaluate and select tools.

We formalize a probabilistic measurement framework built on four constructs, each derived from a distinct failure mode of existing measurement approaches: **inclusion** ($P(v \mid q, m)$) addresses the stochasticity of LLM outputs; **stability** (variance of $\mathbb{1}_{v \in r_i}$, reported as mean inclusion frequency) addresses the unreliability of single-observation measurement; **influence** ($P(\text{role} \mid v, q)$) addresses the aggregation failure of mention counts; and **problem-space coverage** ($P(v \mid ps)$) addresses the conditionality of vendor visibility on buyer context.

The central claim of this framework: *in AI-driven discovery, visibility is no longer a ranked*

position but a probability distribution over generated answers.

Across 81 GPT-4o responses to 27 queries spanning 9 problem areas, we observe: (1) a mean Jaccard overlap of 0.03 between Google and GPT-4o discovery outputs, with 19 of 27 queries at zero overlap; (2) a fragmented vendor landscape of 80 distinct vendors led by Pipedrive (12/27, 0.44)—no vendor achieves universal coverage; (3) role divergence between high-inclusion vendors (in this dataset: Pipedrive Recommended in 84% of instances; Salesforce in 0%); (4) strong problem-space anchoring of context-specific vendors that aggregate inclusion rates obscure; and (5) significant stability variation requiring multi-run measurement to distinguish reliable signal from sampling noise.

Iteration 2 will add Claude as a second LLM, implement two-annotator inter-rater reliability (Cohen's κ), add bootstrap confidence intervals, and conduct a controlled system prompt comparison.

This work establishes a shift in measurement theory: from deterministic ranking-based metrics to probabilistic distribution-based metrics for AI-driven discovery. That shift is not incremental—it is the necessary foundation for any measurement system that takes the stochastic nature of LLM outputs seriously. Any system that reports AI visibility as a single number is fundamentally mis-specified.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, pp. 5–16. ACM. DOI: 10.1145/3637528. arXiv:2311.09735.

Kumar, A., & Lakkaraju, H. (2024). Manipulating Large Language Models to Increase Product Visibility. *arXiv preprint arXiv:2404.07981*. https://arxiv.org/abs/2404.07981

Spatharioti, S.E., Rothschild, D.M., Goldstein, D.G., & Hofman, J.M. (2023). Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. *arXiv preprint arXiv:2307.03744*. https://arxiv.org/abs/2307.03744

Zhang, P., Ye, Q., Peng, Z., Garimella, K., & Tyson, G. (2025). Source Coverage and Citation Bias in LLM-based vs. Traditional Search Engines. *arXiv preprint arXiv:2512.09483*. https://arxiv.org/abs/2512.09483

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.Y., & Wen, J.R. (2023). A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*. https://arxiv.org/abs/2303.18223

SerpAPI. (2026). Google Search API documentation. https://serpapi.com/search-api

OpenAI. (2026). GPT-4o model documentation. https://platform.openai.com/docs

A. System Prompt

The following system prompt was applied uniformly to all GPT-4o API calls:

```
"You are a senior B2B SaaS sales advisor helping a VP of Sales or RevOps leader
at an 80-200 person company.  Answer their question directly and practically.
When relevant, mention specific CRM vendors or tools by name.  Be specific about
trade-offs.  Do not hedge excessively.  Respond in plain prose, 200-400 words."
```

Model: `gpt-4o`. Temperature: 0. Search enabled via `web_search_preview` tool (forced active). Max tokens: 1000. No system message variation across runs.

B. Query Provenance

All queries were derived from real buyer discussions in online communities. Selected source threads:

Query IDs	Source URL	Subreddit
Q1-P, Q1-A	reddit.com/r/sales/comments/1r6ffw2/	r/sales
Q1-B, Q4-P, Q5-P	reddit.com/r/AI_Agents/comments/1s7ghgm/	r/AI_Agents
Q2-P, Q2-B, Q8-B	reddit.com/r/CRMSoftware/comments/1rz38zh/	r/CRMSoftware
Q2-A, Q3-P, Q7-P, Q7-B	reddit.com/r/CRMSoftware/comments/1qgbhkb/	r/CRMSoftware
Q3-A	reddit.com/r/CRMSoftware/comments/1rdg6m5/	r/CRMSoftware
Q3-B, Q7-A, Q9-P, Q9-A, Q9-B	reddit.com/r/sales/comments/16ccmth/	r/sales
Q4-A, Q5-A, Q5-B	reddit.com/r/CRM/comments/1qgx475/	r/CRM
Q6-P, Q6-A	reddit.com/r/CRM/comments/1rcdv58/	r/CRM
Q6-B	reddit.com/r/b2b_sales/comments/1pxk15r/	r/b2b_sales
Q8-P, Q8-A	reddit.com/r/CRMSoftware/comments/1sagtck/	r/CRMSoftware

C. Vendor Inclusion Summary Table

Table 5. Vendor inclusion across all 27 queries and 9 problem areas (GPT-4o, search-enabled, temperature 0). 80 distinct vendors appeared; table shows top 25 by query count.

Vendor	Queries (of 27)	Rate	Areas (of 9)	Dominant role
Pipedrive	12	0.44	8	Recommended
Zoho CRM	10	0.37	6	Recommended
Salesforce	9	0.33	6	Mentioned
HubSpot CRM	9	0.33	4	Recommended
HubSpot	7	0.26	5	Mentioned
Freshsales	7	0.26	5	Recommended
Salesmate	5	0.19	3	Recommended
Close CRM	4	0.15	2	Recommended
Salesflare	3	0.11	3	Recommended
Capsule CRM	3	0.11	3	Recommended
Insightly	3	0.11	3	Recommended
HubSpot Sales Hub	3	0.11	2	Recommended
Apollo.io	3	0.11	2	Recommended
SalesDirector.ai	2	0.07	2	Recommended
Arahi AI	2	0.07	2	Recommended
Rattle	2	0.07	2	Recommended
OnePageCRM	2	0.07	2	Recommended
Zendesk Sell	2	0.07	2	Recommended
Insightly CRM	2	0.07	2	Recommended
Microsoft Dynamics 365	2	0.07	2	Recommended
Salesmate CRM	2	0.07	2	Recommended
Salesloft	2	0.07	2	Mentioned
58 additional vendors each appeared in 1 query (inclusion rate 0.04)				