METHODOLOGICAL STUDY · FA-MET-01

How to Measure What an AI Recommends

How many queries it takes to tell an isolated observation from a real pattern

---

Base: three experiments · 7,000 runs · ChatGPT interface · 2 market categories

Version 1.0 · April 30, 2026 · Translated from the Portuguese original

Rodrigo Beckmann
Co-founder and CTO
First Answer

1. Executive summary

Ask an AI once which brand is best in a category and you get a sample, not a picture. This study measures how many queries it takes to turn samples into a reliable picture, through three experiments in the ChatGPT interface: 1,000 runs in a concentrated market, 1,000 in a fragmented market, and a 5,000-run series in the same fragmented market, for a total of 7,000 runs.

1. **There is a stable pattern behind the variation.** At scale, each brand's presence converges to a measurable value. In the concentrated market, two brands appear in 99.9% of responses and five form a core above 50%.
2. **The shape of the pattern varies across markets.** In the fragmented market, the leader appears in 70.3% of responses, and the universe reaches 144 brands in 1,000 runs, against 38 in the concentrated market.
3. **Precision follows classical statistics.** With 1,000 runs, each brand's presence is measured with a margin of up to ±3.1 percentage points. In a 5,000-run series, none of the top ten brands moved outside that margin between N = 1,000 and N = 5,000: the largest shift was 2.9 points.
4. **The tail requires a different scale.** In the fragmented market, the brand universe is still growing at 5,000 runs, ever more slowly. Of the brands seen only once or twice in the first 1,000 runs, 29% did not return in the next 4,000.

> **Why it matters**
>
> A screenshot does not prove a pattern. Rounds of 1,000 runs measure the visible hierarchy within ±3 points and detect changes above roughly 4 points between rounds. ±1-point precision takes close to 10,000 runs, and mapping the tail of a fragmented market takes thousands. Section 4.4 gives the recommended N by objective.

2. The question

ChatGPT alone receives about 2.5 billion prompts a day (TechCrunch, 2025), and each person walks away feeling they have seen what the AI thinks. Almost nobody looks at the aggregate. Asked again, in a new conversation, the same question could bring other brands in another order, because these models generate text by sampling among plausible options.

Picture someone who asks and sees Brand X in the response. Brand X might appear in 99% of the responses the AI would give, in 25%, or in 0.8%. The observable scene is identical in all three cases, and all three produce the same claim on social media, but the meaning is radically different.

> *How can we separate, with statistical rigor, an isolated observation from a real pattern in the behavior of conversational AIs, and how many runs does that separation require to be reliable?*

What is already known

Aggarwal et al. (2024) formalized Generative Engine Optimization (GEO) and proposed metrics for the visibility of content in generated responses, evaluated over a broad set of queries. Their focus is what content creators can do to appear, not how much a visibility measurement must collect to be reliable.

Fishkin and O'Donnell (2026) measured the consistency of brand recommendations: 12 prompts run nearly 3,000 times across ChatGPT, Claude and Google's AI. The chance of two responses returning the same list was below 1 in 100, but the frequency with which each brand appears proved far more stable than its position, and the authors recommend measuring presence rather than rank. The study used 60 to 100 runs per prompt, was carried out manually by volunteers, and left open how many runs are needed for measured visibility to be reliable.

This study starts where that work stops: it takes presence as the metric and asks how precisely it is measured at each N, how the brand universe grows with collection, and what changes between markets with different structures.

Hypothesis and criterion

Hypothesis tested: each brand's presence, measured over independent runs, converges to a stable value and behaves as a binomial proportion, with the precision theory predicts.

Refutation criterion: the hypothesis fails if, when the sample is multiplied, the presence of the leading brands moves beyond the 95% margin predicted for the smaller sample.

Experiment 1 tests the question in a concentrated market. Experiment 2 asks whether the pattern holds in a fragmented market. Experiment 3 revisits the fragmented market with a 5,000-run series, to find out whether the brand universe is open or merely large and whether the 1,000-run hierarchy holds.

3. Method

Environment. All three experiments were collected through the ChatGPT interface, in the browser, by automation (web scraping). This is the environment in which consumers ask the question. This study's structural conclusions do not depend on the environment; the brand composition measured here holds for the interface.

Date, language and search. All runs took place on April 27, 2026, in Brazilian Portuguese, with web search enabled.

Question and isolation. Each experiment used a single, fixed question, written the way a consumer would type it, with no brands as anchors and no format request. Each run opened a new conversation with no history, so that one response could not influence the next.

Extraction and normalization. From each response we extracted every brand mentioned, automatically, with human review by sampling. Every brand mentioned counted as present, with no distinction between primary and secondary mentions. Product lines, sub-brands and models were attributed to the parent brand; retailers, marketplaces and generic terms were filtered out, and when in doubt the term was kept.

Anonymization. Brands are identified by letters, in order of presence at N = 1,000 in each experiment. Letters do not change once assigned and are not comparable across experiments: Brand A in Experiment 1 is not Brand A in Experiment 2.

Statistics. Runs in new conversations are independent samples, so a brand’s presence follows the binomial distribution. 95% intervals use the Wilson method (Wilson, 1927). Brands with presence below 0.3% in 1,000 runs are treated together, as the rare tail. The full study sheet is in Appendix A.

4. Findings

4.1 Experiment 1 — the deterministic core in a concentrated market

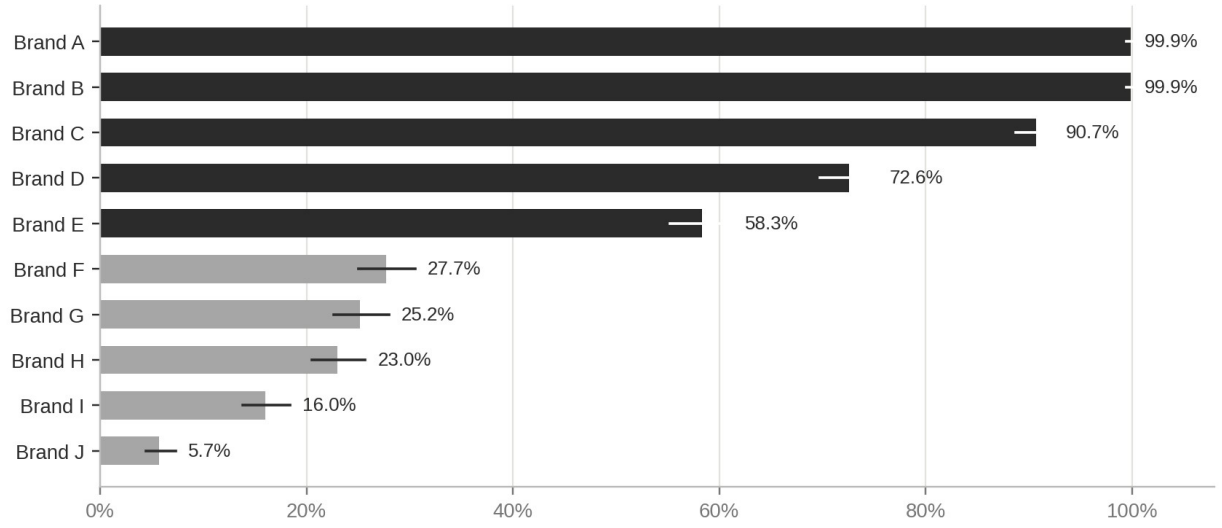
The first experiment's category has few relevant players and brands most consumers recognize. It is the most favorable setting for a pattern to emerge: if it did not show up here, it would hardly show up anywhere. There were 1,000 identical runs, mentioning 38 distinct brands.

Finding 1.1 — Presence is highly unequal

Two brands, A and B, appeared in 999 of the 1,000 responses. Right below come Brand C (90.7%), D (72.6%) and E (58.3%). These five brands appear in more than half of the responses and form the category's deterministic core.

Five brands appear in more than half of the responses

Presence of the 10 most frequent brands. Dark: core (>50%). Line: 95% interval (Wilson).



Source: FA-MET-01 v1.0 · Exp. 1, concentrated market · 1,000 runs

Figure 1 — Presence of the 10 most frequent brands in Experiment 1.

Table 1 — Presence of the 10 most frequent brands in Experiment 1

Brand	Appearances	Presence	95% CI
Brand A	999	99.9%	99.4% to 100.0%
Brand B	999	99.9%	99.4% to 100.0%
Brand C	907	90.7%	88.7% to 92.3%
Brand D	726	72.6%	69.8% to 75.3%
Brand E	583	58.3%	55.2% to 61.3%
Brand F	277	27.7%	25.0% to 30.6%
Brand G	252	25.2%	22.6% to 28.0%
Brand H	230	23.0%	20.5% to 25.7%
Brand I	160	16.0%	13.9% to 18.4%
Brand J	57	5.7%	4.4% to 7.3%

Base: 1,000 responses. Wilson interval.

Below the core sits a middle band, from 16% to 28% (Brands F to I), that appears often but without guarantee. And there is a short tail: Brand J appears in 5.7% of responses.

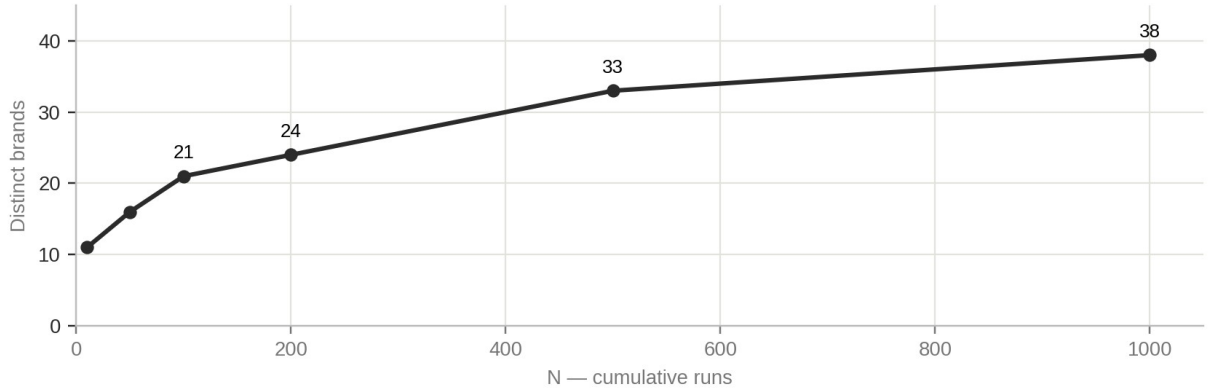
This finding does not show that every market has a core: Experiment 2 shows a market with no brand above 71%.

Finding 1.2 — The tail is short and saturates quickly

The first 10 responses had already surfaced 11 brands; 100 responses, 21; 500, 33. The next 500 responses added only 5, closing at 38. After N = 600, fewer than one new brand appears per 100 responses. Of the 38 brands, 14 appeared only once or twice.

The brand universe saturates near 40 in the concentrated market

Distinct brands observed as N grows. Measured points; the line only connects them.



Source: FA-MET-01 v1.0 · Exp. 1, concentrated market · 1,000 runs

Figure 2 — Distinct brands observed in Experiment 1 as N grows.

The model does not draw brands from an open universe: it works with a relatively closed vocabulary of candidates for the question, and each response is a selection from that vocabulary with very unequal weights. For brands, being in that vocabulary has become a prerequisite for appearing. A brand that does not surface in the first few hundred runs is unlikely to surface with more volume.

Finding 1.3 — Apparent randomness is a lack of samples

A coin tossed 10 times can land heads 7 times, 70%. Tossed 1,000 times, it lands very close to 50%. That is the law of large numbers, and it describes what happens with responses: each run is a sample from a distribution that exists inside the model, and each brand's observed frequency approaches its true value as samples accumulate.

The data show this happening. Brands A and B passed 99% before N = 100. Brand C swung between 82% and 92% over the first 50 responses and settled around 91% from N = 300 on. Middle-band brands fluctuated until N = 200 and then settled. Between the N = 100–200 window and the N = 900–1,000 window, the standard deviation of the cumulative frequency fell by 60% to 95% for the top ten brands.

4.2 Experiment 2 — what changes in a fragmented market

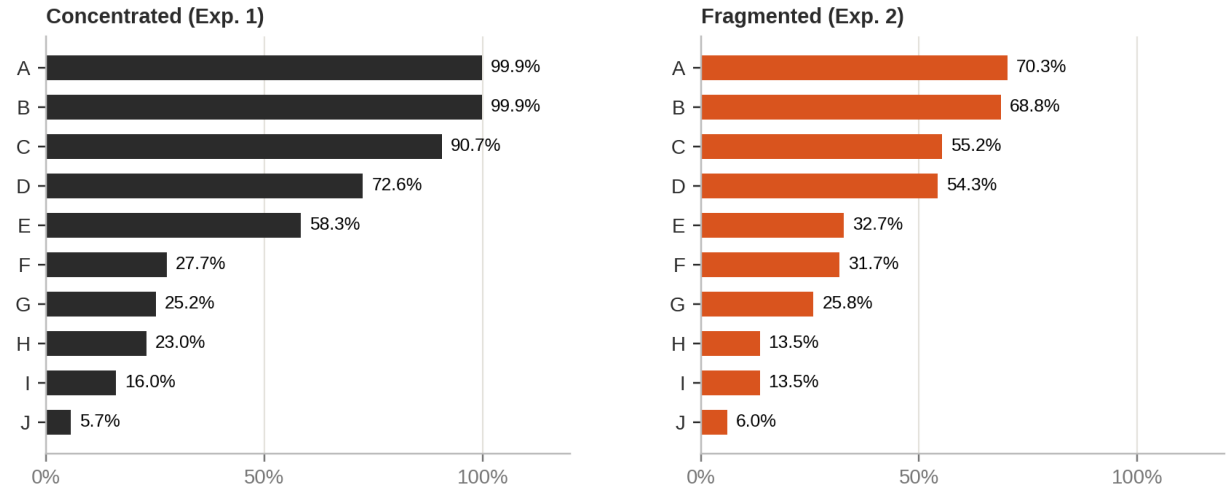
The second experiment repeated the method in a category with the opposite profile: dozens of legitimate brands, domestic and international, competing with no absolute leaders. There were 1,000 runs, mentioning 144 distinct brands, almost four times Experiment 1.

Finding 2.1 — The shape of the distribution changes; the backbone repeats

In the fragmented market no brand comes close to 99.9%: the leader appears in 70.3% of responses and the runner-up in 68.8%. The drop-off is different too. The concentrated market has a step between the core and the middle band (the fifth brand appears in 58%, the sixth in 28%). The fragmented market has a ramp: the first four go from 70% to 54%, and the next three sit between 25% and 33%.

In the fragmented market, no brand exceeds 71%

Presence of the 10 most frequent brands in each experiment (1,000 runs each). Letters independent per experiment.



Source: FA-MET-01 v1.0 · Exp. 1 and Exp. 2

Figure 3 — Presence of the 10 most frequent brands in each experiment.

Table 2 — Structural comparison of Experiments 1 and 2

Metric	Concentrated	Fragmented
Distinct brands in 1,000 runs	38	144
Presence of brand #1	99.9%	70.3%
Presence of brand #2	99.9%	68.8%
Presence of brand #3	90.7%	55.2%
Presence of brand #5	58.3%	32.7%
Brands with presence ≥ 50%	5	4
Brands with presence ≥ 25%	7	7
Brands with presence ≥ 10%	9	9
Brands seen only once or twice	14	119
Top 5 — % of mention volume	78%	70%
Top 10 — % of mention volume	96%	93%

Base: 1,000 responses per experiment.

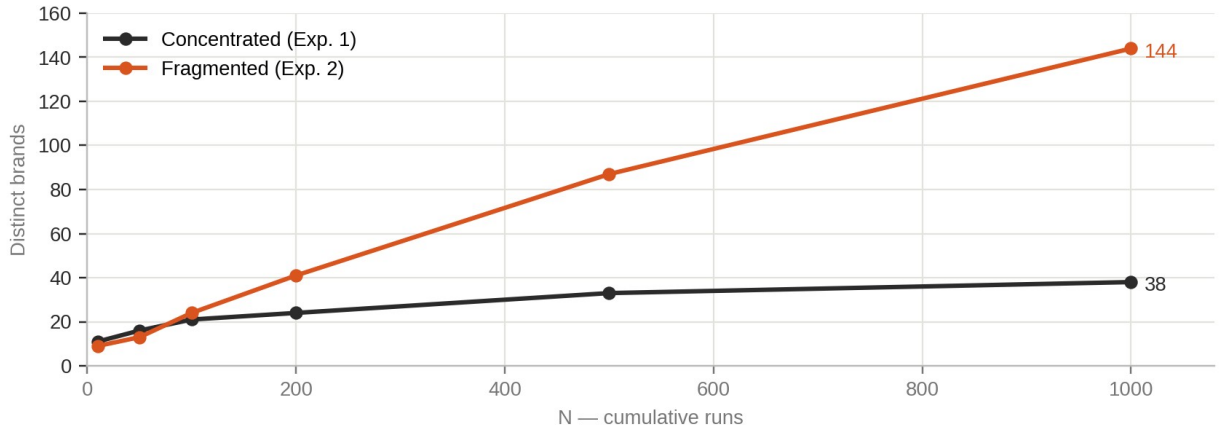
The fragmented market has 3.8 times as many brands, yet the number of brands above 25% (seven) and above 10% (nine) is identical in both. A structural backbone repeats: the hierarchy is more spread out in the fragmented market, in lower presence bands, with the same number of brands at each step.

Finding 2.2 — The long tail is only long in the fragmented market

At N = 100 the two experiments had seen a similar number of brands (21 and 24). From N = 200 they diverge: the concentrated market has 24 brands, close to saturation, while the fragmented one passes 40 and keeps growing to 144 at N = 1,000.

The tail is only long in the fragmented market

Distinct brands observed as N grows, in both experiments. Measured points.



Source: FA-MET-01 v1.0 · Exp. 1 and Exp. 2

Figure 4 — Distinct brands observed as N grows, in both experiments.

Table 3 — Distinct brands observed by N

N	Concentrated	Fragmented
10	11	9
50	16	13
100	21	24
200	24	41
500	33	87
1,000	38	144

Base: 1,000 responses per experiment, in collection order.

There are at least two vocabulary regimes. In the concentrated market it is closed, with visible saturation before N = 500. In the fragmented market it is broad and keeps growing beyond 1,000. Experiment 3 tests how far.

Finding 2.3 — The volume captured by the top is similar in both markets

Despite the difference at the top, the five most frequent brands account for 78% of mentions in the concentrated market and 70% in the fragmented one; the top ten, 96% and 93% (Table 2). For anyone trying to understand who dominates and to map the competition, the top ten brands capture almost the same volume in both kinds of market. The real difference lies in the tail.

4.3 Experiment 3 — a 5,000-run series

Experiment 2 left a question open: at 1,000 runs, the tail of the fragmented market was still growing. Is the universe open, or just large? To answer, we ran a series of 5,000 identical runs, with the same question and the same method as Experiment 2.

> *Does the hierarchy measured at 1,000 runs hold at 5,000? And is the brand universe of the fragmented market open, or merely large?*

Finding 3.1 — The 1,000-run hierarchy stayed within the margin

We compared the presence of the top ten brands at N = 1,000 and at N = 5,000. Every shift stayed below 3 points; the largest was Brand D's, from 54.3% to 57.2%. No brand left the 95% interval computed at N = 1,000.

Table 4 — Presence of the top 10 brands at N = 1,000 and N = 5,000

Brand	N = 1,000	95% CI at 1,000	N = 5,000	Δ (pp)	Within CI?
Brand A	70.3%	67.4% to 73.1%	69.4%	−0.9	Yes
Brand B	68.8%	65.9% to 71.6%	68.2%	−0.6	Yes
Brand C	55.2%	52.1% to 58.3%	55.9%	+0.7	Yes
Brand D	54.3%	51.2% to 57.4%	57.2%	+2.9	Yes
Brand E	32.7%	29.9% to 35.7%	31.5%	−1.2	Yes
Brand F	31.7%	28.9% to 34.6%	32.4%	+0.7	Yes
Brand G	25.8%	23.2% to 28.6%	25.3%	−0.5	Yes
Brand H	13.5%	11.5% to 15.8%	13.8%	+0.3	Yes
Brand I	13.5%	11.5% to 15.8%	13.3%	−0.2	Yes
Brand J	6.0%	4.7% to 7.6%	6.1%	+0.1	Yes

Base: 1,000 and 5,000 responses. Letters in order of presence at N = 1,000 in the series, fixed once assigned.

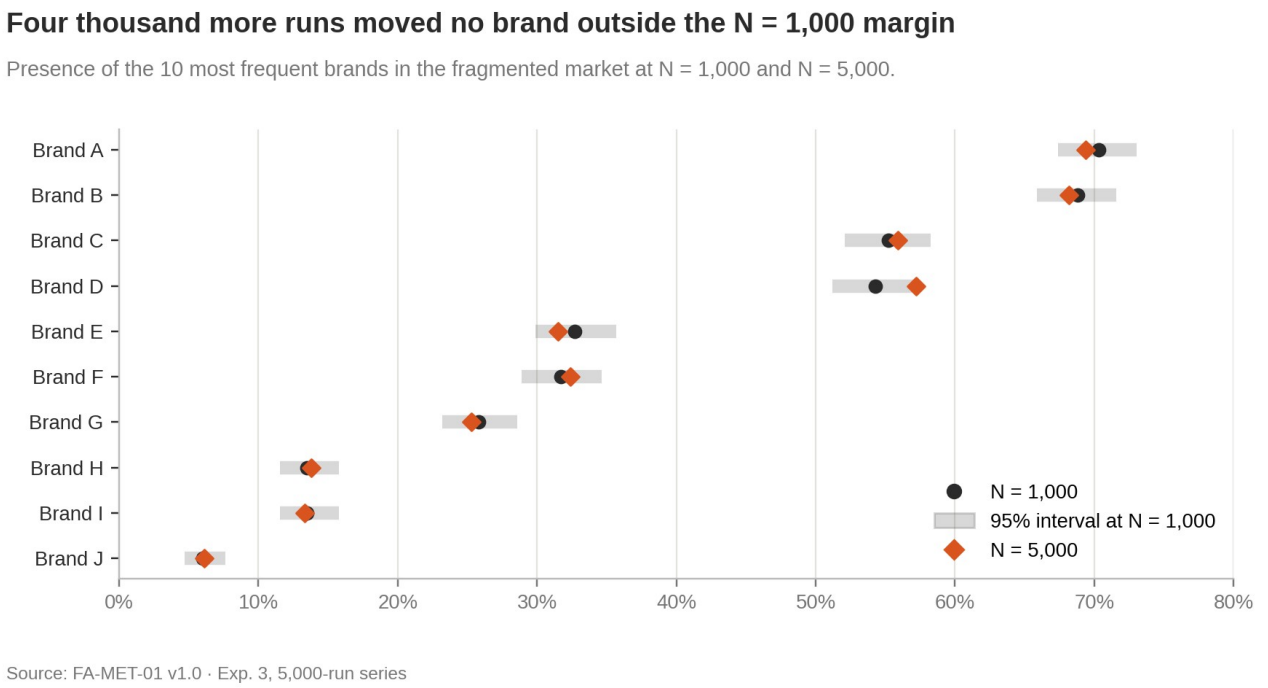


Figure 5 — Presence at N = 1,000 (with 95% interval) and at N = 5,000.

This is exactly what theory predicts. For a brand with 55% presence, the 95% margin at 1,000 runs is ±3.1 points; Brand D moved 2.9. Four thousand additional runs refined the estimates but did not move the hierarchy. The study's hypothesis holds: presence behaves as a binomial proportion, with the precision of classical theory.

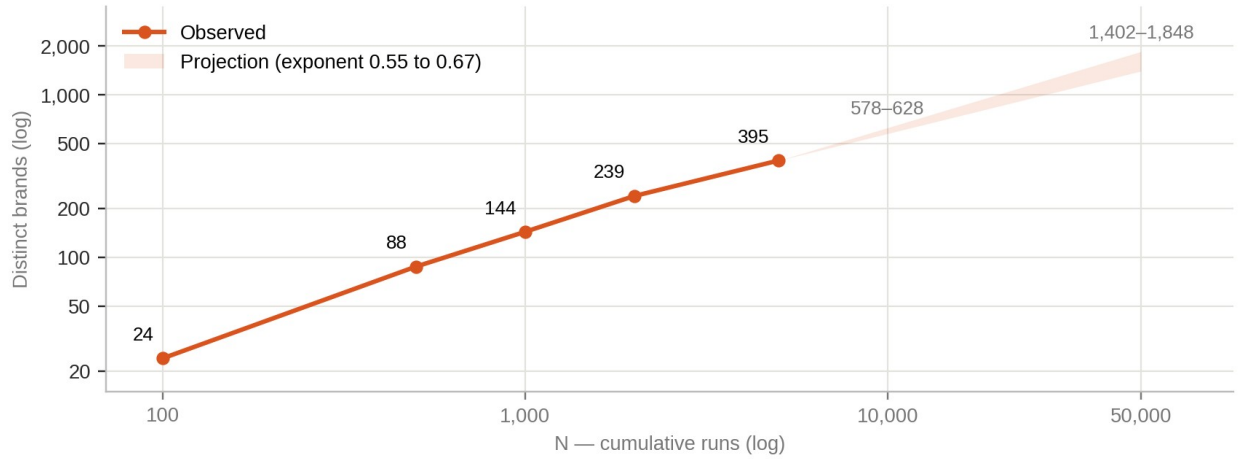
This finding does not show that shifts of a few points between two 1,000-run rounds are noise: comparing two rounds adds up the uncertainty of both (see Section 4.4).

Finding 3.2 — The tail grows, with diminishing returns

The hierarchy was complete at 1,000 runs; the universe was not. In the 5,000-run series, the number of distinct brands went from 144 at N = 1,000 to 395 at N = 5,000, 2.7 times as many for 5 times the effort.

The fragmented market's tail is still growing at 5,000, ever more slowly

Distinct brands observed in the 5,000-run series and power-law projection band.



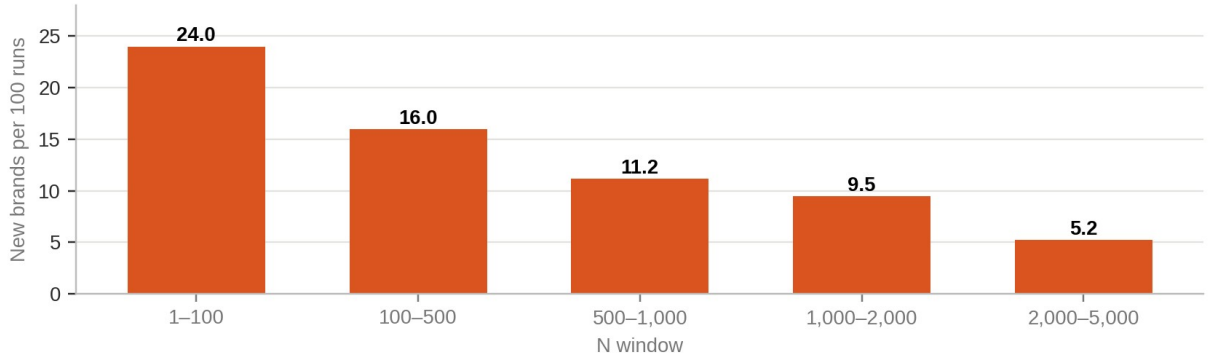
Source: FA-MET-01 v1.0 · Exp. 3. Local exponent fell from 0.73 (1,000→2,000) to 0.55 (2,000→5,000).

Figure 6 — Distinct brands in the 5,000-run series and projection band, on a logarithmic scale.

The discovery rate falls. The first 100 runs surfaced 24 new brands; between 2,000 and 5,000, about 5 per 100 runs.

Every 100 runs surface fewer and fewer new brands

New brands per 100 runs in each window of the 5,000-run series.



Source: FA-MET-01 v1.0 · Exp. 3, 5,000-run series

Figure 7 — New brands per 100 runs in each window of the 5,000-run series.

Vocabulary growth roughly follows a power law, Heaps' law (Heaps, 1978), well known in information retrieval, but the local exponent decreases: it was 0.73 between 1,000 and 2,000 runs and 0.55 between 2,000 and 5,000. Projecting with exponents in that range, 10,000 runs would reveal between 580 and 630 brands, and 50,000 between 1,400 and 1,850. The universe is large and the discovery rate never reaches zero, but each new band of runs yields less.

The projections extrapolate five measured points and serve to size data collection, not to state the size of the model's vocabulary.

Finding 3.3 — Part of a small sample's tail does not hold up

In the first 1,000 runs of the series, 119 brands appeared only once or twice. The next 4,000 runs make it possible to test whether they belong to the model's vocabulary: a real brand with low presence would have four times as many chances to reappear.

71% of the rare brands at N = 1,000 returned in the next 4,000 runs

119 brands seen only once or twice in the first 1,000 runs of the 5,000-run series.



Source: FA-MET-01 v1.0 · Exp. 3, 5,000-run series

Figure 8 — Rare brands from the first 1,000 runs that reappeared in the next 4,000.

Of the 119, 85 (71%) reappeared. The other 34 (29%) did not return. Some may be one-off captures, such as a product model with no parent brand or a poorly extracted entity; others may be real brands with presence so low (around 0.02%) that 4,000 runs are not enough to see them again. The data cannot tell the two apart.

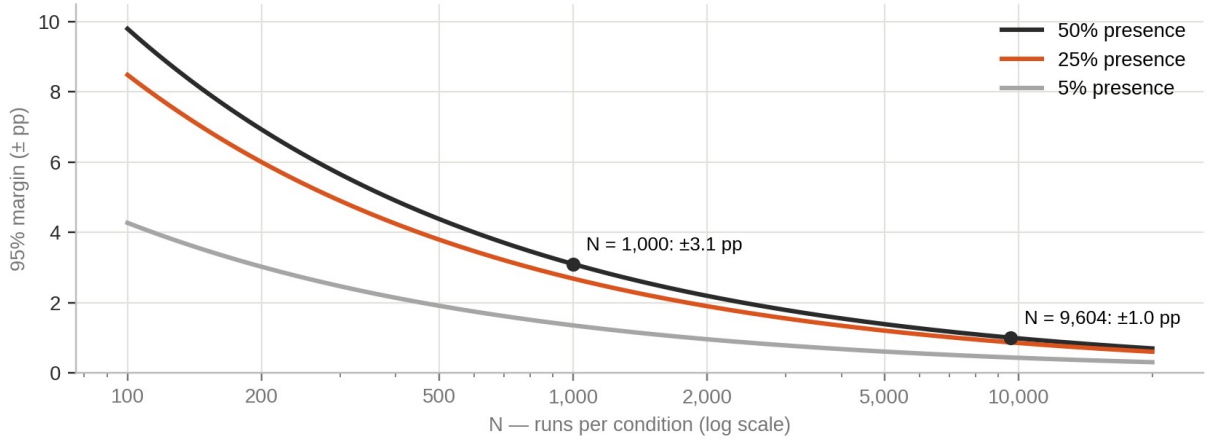
The practical implication does not depend on that distinction: a tail list drawn from 1,000 runs contains about one third unconfirmed brands and should carry that warning.

4.4 How many runs are needed

The precision of a presence measurement depends only on N and on the presence itself. Figure 9 shows the 95% margin as N grows.

At 1,000 runs the margin is ±3 pp; ±1 pp takes ~9,600

Half-width of the 95% interval for a brand's presence, by N and presence level.



FA-MET-01 v1.0 · binomial calculation: 1.96 × √(p(1−p)/N)

Figure 9 — 95% margin of a brand's presence, by N and by presence level.

Table 5 — Recommended N by objective

Objective	N per condition	What it delivers
Initial diagnosis	100 to 200	Regime (concentrated or fragmented) and approximate top, within ±7 to 10 pp

Objective	N per condition	What it delivers
Visible hierarchy	1,000	Presence within ±3 pp; reliable order between brands more than ~6 pp apart
Change between rounds	1,000 per round	Changes above ~4.4 pp (presence near 50%) can be told apart from noise
±1 pp precision	~9,600	Presence within ±1 pp at any level
Tail down to 0.1%	~3,000	95% chance of seeing at least once each brand with 0.1% presence
Tail down to 0.02%	~15,000	Same, for 0.02%

Margins at 95% confidence. Tail rows: $1 - (1 - p)^N \geq 0.95$.

A common line of reasoning leads one to expect more precision than this: since responses in the same category resemble each other, each run would count for several. It does not. Runs in new conversations are independent, and the similarity between responses is already contained in the presence itself. Nor is the N at which the cumulative mean stays within ±1 point of the same series' final value evidence of anything: that metric converges by construction, because the last runs are part of the final value. Experiment 3 confirms classical theory: 1,000 runs give ±3 points, not ±1.

5. What generalizes and what varies

Supported by all three experiments

- **Convergence through repeated sampling.** Each brand's presence stabilizes as N grows, as the law of large numbers predicts, in any category.
- **A measurable, stable hierarchy.** Some brands consistently appear more than others, and multiplying the sample by five kept the order within the margin.
- **Binomial precision.** The 95% margin follows the classical formula: ±3.1 points at 1,000 runs, ±1 point at about 9,600.
- **An unreliable small-sample tail.** Brands seen once or twice in 1,000 runs include a sizeable share that does not hold up with more volume.

Varies by category and needs its own measurement

- **Shape of the top.** Two brands at 99.9% in the concentrated market; none above 71% in the fragmented one.
- **Size of the universe.** 38 against 144 brands in 1,000 runs.
- **Saturation.** The concentrated market saturates before 1,000 runs; the fragmented one is still growing at 5,000.
- **Composition by environment.** The brands below the core may change depending on how the model is accessed.

6. Implications by reader

For brands and marketers

Being in the model's active vocabulary is no longer a technical detail. Core brands enjoy continuous presence at no media cost; brands outside it face structural invisibility. The first question in any monitoring program is whether the category has a core and who is in it, and 100 to 200 runs answer that. The second is the brand's own standing, which takes rounds of 1,000 runs, bearing in mind that only changes above about 4 points between rounds can be told apart from noise.

For researchers and journalists

A single question to an AI is a single coin toss. The brand mentioned may be a choice that appears in 95% of responses or a draw that appears in 1%. Any claim about what an AI recommends needs hundreds of independent samples, ideally 1,000 or more, repeated periodically. Attributing bias on the basis of rare-tail brands calls for caution: part of that tail does not hold up with more volume.

For AI product builders

If a product depends on an AI recommending something consistently, consistency must be measured, not assumed. Some choices are near-deterministic, above 90%; others vary in 30% or more of responses. More specific questions, format instructions and domain constraints reduce variation, but the first step is to measure the current level with this study's protocol.

7. Limitations

- **One environment.** All three experiments were collected through the ChatGPT interface. The statistical regularity should hold for other models and access environments; the brand composition not necessarily.
- **Model in the interface.** The interface does not allow the answering model to be fixed or recorded. We cannot guarantee that every run was served by the same model.
- **Two categories.** The specific figures (five-brand core, 29% unconfirmed tail, discovery-curve exponent) reflect these categories.
- **Temporal validity.** Model updates may change the ranking. Conclusions about the shape of the curves and about precision tend to be more stable than brand-level figures.
- **One question per experiment.** Other phrasings of the same intent may produce different distributions. That effect was not measured here.
- **One date.** All data were collected on April 27, 2026.
- **Transparency.** The model, the question wording and the categories are not disclosed, and the dataset is not public. The figures can be checked against one another, but not recomputed by third parties from the same inputs.
- **Normalization.** Filtering non-brand terms affects absolute figures, especially in the tail. The list is conservative: when in doubt, the term was kept.

8. Reproducibility

What transfers is the protocol: number of runs, a new conversation per run, a fixed question, binary presence per response, parent-brand normalization and binomial intervals. With it, any team can repeat the design on a model and category of its choice and compare the shape of the curves. Replications on other models, other categories, at other times and via the API are the ones that most refine what generalizes.

9. 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 2024)*. arXiv:2311.09735. https://doi.org/10.48550/arXiv.2311.09735

Fishkin, R.; O'Donnell, P. (2026). NEW Research: AIs are highly inconsistent when recommending brands or products; marketers should take care when tracking AI visibility. *SparkToro*, January 2026. https://sparktoro.com/blog/new-research-ais-are-highly-inconsistent-when-recommending-brands-or-products-marketers-should-take-care-when-tracking-ai-visibility/

Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. New York: Academic Press.

TechCrunch (2025). ChatGPT users send 2.5 billion prompts a day. July 21, 2025. https://techcrunch.com/2025/07/21/chatgpt-users-send-2-5-billion-prompts-a-day/

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. https://doi.org/10.2307/2276774

Appendix A — Study sheet

Field	Record
Code and version	FA-MET-01 · v1.0
Question	One per experiment, fixed, phrased as a purchase intent, with no brand anchors. Wording not disclosed
Categories	Exp. 1: concentrated market. Exp. 2 and 3: fragmented market (same category). Not disclosed
Language	Brazilian Portuguese
Environment	ChatGPT interface in the browser, by automation (web scraping), with web search enabled
Collection date	April 27, 2026
Isolation	New conversation per run, no history
N	Exp. 1: 1,000. Exp. 2: 1,000. Exp. 3: 5,000. Total: 7,000 runs
Unit of analysis	Response × brand present (binary presence per response)
Normalization	Parent-brand consolidation; retailers, marketplaces and generic terms filtered out; when in doubt, the term was kept
Anonymization	Letters in order of presence at N = 1,000 in each experiment, fixed once assigned
Statistics	95% Wilson intervals; individual analysis threshold at 0.3% presence

Appendix B — Version history

Date	Version	Changes
April 30, 2026	1.0	First published version. English translation of the Portuguese original, with the same figures and code.