ESTUDO METODOLÓGICO · FA-MET-01

Como medir o que uma IA recomenda

Quantas perguntas são necessárias para separar observação isolada de padrão real

---

Base: três experimentos · 7.000 chamadas · interface do ChatGPT · 2 categorias de mercado

Versão 1.0 · 30 de abril de 2026

Rodrigo Beckmann
Cofundador e CTO
First Answer

1. Resumo executivo

Quem pergunta uma vez a uma IA qual a melhor marca de uma categoria recebe uma amostra, não um retrato. Este estudo mede quantas perguntas são necessárias para transformar amostras em retrato confiável, em três experimentos na interface do ChatGPT: 1.000 chamadas num mercado consolidado, 1.000 num mercado fragmentado e uma série de 5.000 no mesmo mercado fragmentado, totalizando 7.000 chamadas.

1. **Existe padrão estável por trás da variação.** Em escala, a presença de cada marca converge para um valor mensurável. No mercado consolidado, duas marcas aparecem em 99,9% das respostas e cinco formam um núcleo acima de 50%.
2. **A forma do padrão varia entre mercados.** No mercado fragmentado, a líder aparece em 70,3% das respostas, e o universo chega a 144 marcas em 1.000 chamadas, contra 38 no consolidado.
3. **A precisão segue a estatística clássica.** Com 1.000 chamadas, a presença de cada marca é medida com margem de até ±3,1 pontos percentuais. Numa série de 5.000 chamadas, nenhuma das dez marcas principais saiu dessa margem entre N = 1.000 e N = 5.000: a maior variação foi de 2,9 pontos.
4. **A cauda exige outra escala.** No mercado fragmentado, o universo de marcas ainda cresce em 5.000 chamadas, cada vez mais devagar. Das marcas vistas só uma ou duas vezes nas primeiras 1.000 chamadas, 29% não voltaram nas 4.000 seguintes.

> **O que isso muda**
>
> Um print não prova padrão. 1.000 chamadas por rodada medem a hierarquia visível com ±3 pontos e detectam mudanças acima de cerca de 4 pontos entre rodadas. Precisão de ±1 ponto exige perto de 10.000 chamadas, e mapear a cauda de um mercado fragmentado exige milhares. A Seção 4.4 traz o N recomendado por objetivo.

2. A pergunta

Só o ChatGPT recebe cerca de 2,5 bilhões de perguntas por dia (TechCrunch, 2025), e cada pessoa sai com a impressão de ter visto o que a IA pensa. Quase ninguém olha o agregado. Se a mesma pergunta fosse feita de novo, numa conversa nova, poderia trazer outras marcas em outra ordem, porque esses modelos geram texto amostrando entre opções plausíveis.

Considere alguém que pergunta e vê a Marca X na resposta. A Marca X pode aparecer em 99% das respostas que a IA daria, em 25% ou em 0,8%. A cena observável é a mesma nos três casos, e as três geram a mesma afirmação nas redes sociais, mas o significado é radicalmente diferente.

> *Como separar, com rigor estatístico, observação isolada de padrão real no comportamento de IAs conversacionais, e quantas chamadas são necessárias para essa separação ser confiável?*

O que já se sabe

Aggarwal et al. (2024) formalizaram a otimização para motores generativos (GEO) e propuseram métricas de visibilidade de conteúdo nas respostas, avaliadas num conjunto amplo de consultas. O foco do trabalho é o que o criador de conteúdo pode fazer para aparecer, não quanto uma medição de visibilidade precisa coletar para ser confiável.

Fishkin e O'Donnell (2026) mediram a consistência das recomendações de marca: 12 prompts executados quase 3.000 vezes em ChatGPT, Claude e IA do Google. A chance de duas respostas trazerem a mesma lista foi menor que 1 em 100, mas a frequência com que cada marca aparece se mostrou bem mais estável que a posição, e os autores recomendam medir presença em vez de ranking. O estudo usou de 60 a 100 execuções por prompt, foi feito por voluntários, manualmente, e deixou em aberto quantas execuções são necessárias para que a visibilidade medida seja confiável.

Este estudo parte do ponto em que esse trabalho para: aceita a presença como métrica e pergunta com que precisão ela é medida para cada N, como o universo de marcas cresce com a coleta e o que muda entre mercados de estrutura diferente.

Hipótese e critério

Hipótese testada: a presença de cada marca, medida por chamadas independentes, converge para um valor estável e se comporta como uma proporção binomial, com a precisão prevista pela teoria.

Critério de refutação: a hipótese cai se, ao multiplicar a amostra, a presença das marcas principais variar além da margem de 95% prevista para a amostra menor.

O Experimento 1 testa a pergunta num mercado consolidado. O Experimento 2 pergunta se o padrão se repete num mercado fragmentado. O Experimento 3 repete o mercado fragmentado numa série de 5.000 chamadas, para saber se o universo de marcas é aberto ou apenas grande e se a hierarquia de 1.000 chamadas se mantém.

3. Método

Ambiente. Os três experimentos foram coletados na interface do ChatGPT, no navegador, por automação (web scraping). É o ambiente em que o consumidor faz a pergunta. As conclusões estruturais deste estudo não dependem do ambiente; a composição das marcas medida aqui vale para a interface.

Data, idioma e busca. Todas as chamadas foram feitas em 27 de abril de 2026, em português do Brasil, com a busca na web habilitada.

Pergunta e isolamento. Cada experimento usou uma pergunta única e fixa, escrita como um consumidor digitaria, sem marcas como âncora nem pedido de formato. Cada chamada abriu uma conversa nova, sem histórico, para que uma resposta não influenciasse a seguinte.

Extração e normalização. De cada resposta extraímos todas as marcas citadas, de forma automatizada, com revisão humana por amostragem. Toda marca citada contou como presente, sem distinguir menção principal de secundária. Linhas, submarcas e modelos foram atribuídos à marca-mãe; varejistas, marketplaces e termos genéricos foram filtrados, e em caso de dúvida o termo foi mantido.

Anonimização. As marcas são identificadas por letras, pela ordem de presença em N = 1.000 de cada experimento. As letras não mudam depois de atribuídas e não se comparam entre experimentos: a Marca A do Experimento 1 não é a Marca A do Experimento 2.

Estatística. Chamadas em conversas novas são amostras independentes, então a presença de uma marca segue a distribuição binomial. Os intervalos de 95% usam o método de Wilson (Wilson, 1927). Marcas com presença abaixo de 0,3% em 1.000 chamadas são tratadas em conjunto, como cauda rara. A ficha técnica completa está no Anexo A.

4. Achados

4.1 Experimento 1 — o núcleo determinístico em mercado consolidado

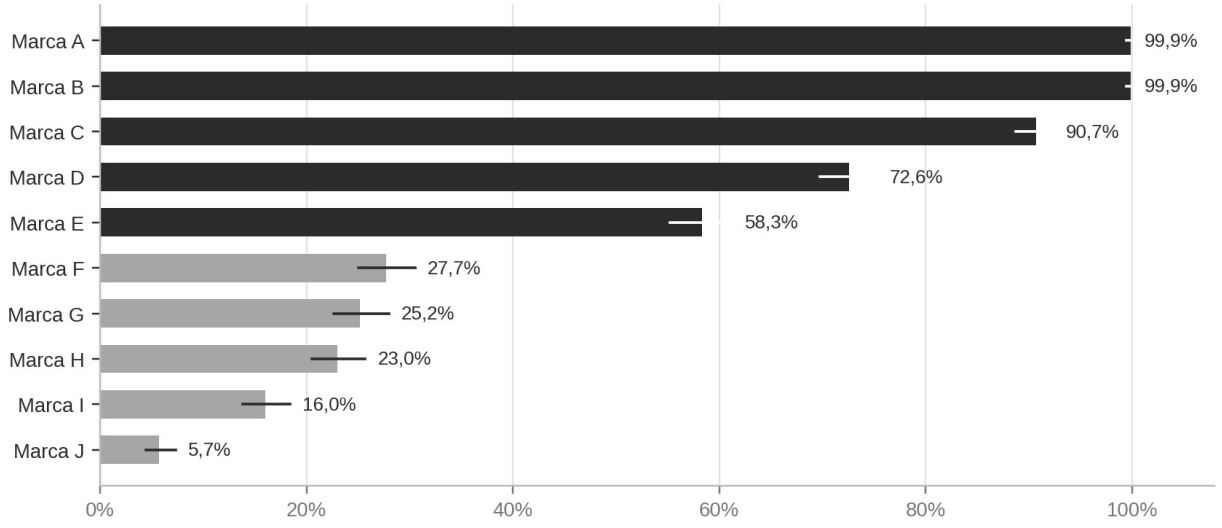
A categoria do primeiro experimento tem poucos players relevantes e marcas que a maior parte dos consumidores reconhece. É o cenário mais favorável para um padrão aparecer: se não aparecesse aqui, dificilmente apareceria em outro lugar. Foram 1.000 chamadas idênticas, que citaram 38 marcas distintas.

Achado 1.1 — A presença é fortemente desigual

Duas marcas, A e B, apareceram em 999 das 1.000 respostas. Logo abaixo vêm a Marca C (90,7%), a D (72,6%) e a E (58,3%). Essas cinco marcas aparecem em mais da metade das respostas e formam o núcleo determinístico da categoria.

Cinco marcas aparecem em mais da metade das respostas

Presença das 10 marcas mais frequentes. Escuro: núcleo (>50%). Traço: intervalo de 95% (Wilson).



Fonte: FA-MET-01 v1.0 · Exp. 1, mercado consolidado · 1.000 chamadas

Figura 1 — Presença das 10 marcas mais frequentes no Experimento 1.

Tabela 1 — Presença das 10 marcas mais frequentes no Experimento 1

Marca	Aparições	Presença	IC 95%
Marca A	999	99,9%	99,4% a 100,0%
Marca B	999	99,9%	99,4% a 100,0%
Marca C	907	90,7%	88,7% a 92,3%
Marca D	726	72,6%	69,8% a 75,3%
Marca E	583	58,3%	55,2% a 61,3%
Marca F	277	27,7%	25,0% a 30,6%
Marca G	252	25,2%	22,6% a 28,0%
Marca H	230	23,0%	20,5% a 25,7%
Marca I	160	16,0%	13,9% a 18,4%
Marca J	57	5,7%	4,4% a 7,3%

Base: 1.000 respostas. Intervalo de Wilson.

Abaixo do núcleo há uma faixa média, de 16% a 28% (Marcas F a I), que aparece com frequência mas sem garantia. E há uma cauda curta: a Marca J aparece em 5,7% das respostas.

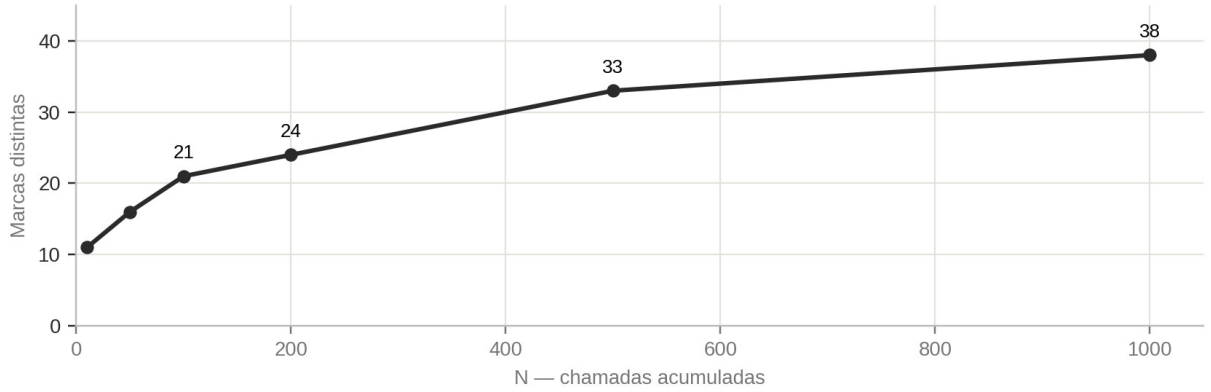
Este achado não permite afirmar que todo mercado tem núcleo: o Experimento 2 mostra um mercado sem nenhuma marca acima de 71%.

Achado 1.2 — A cauda é curta e satura rápido

Nas primeiras 10 respostas já tinham aparecido 11 marcas; em 100, 21; em 500, 33. As 500 respostas seguintes acrescentaram só 5, fechando em 38. Depois de N = 600, surge menos de uma marca nova a cada 100 respostas. Das 38 marcas, 14 apareceram só uma ou duas vezes.

O universo de marcas satura perto de 40 no mercado consolidado

Marcas distintas observadas conforme N cresce. Pontos medidos; a linha só os liga.



Fonte: FA-MET-01 v1.0 · Exp. 1, mercado consolidado · 1.000 chamadas

Figura 2 — Marcas distintas observadas no Experimento 1 conforme N cresce.

O modelo não busca marcas num universo aberto: trabalha com um vocabulário relativamente fechado de candidatas para a pergunta, e cada resposta é uma seleção desse vocabulário com pesos muito desiguais. Para as marcas, estar nesse vocabulário virou pré-requisito para aparecer. Uma marca que não surge nas primeiras centenas de chamadas dificilmente surgirá com mais volume.

Achado 1.3 — A aleatoriedade aparente é falta de amostras

Uma moeda jogada 10 vezes pode dar 7 caras, 70%. Jogada 1.000 vezes, dá muito perto de 50%. É a lei dos grandes números, e ela descreve o que acontece com as respostas: cada chamada é uma amostra de uma distribuição que existe dentro do modelo, e a frequência observada de cada marca se aproxima do valor verdadeiro conforme as amostras se acumulam.

Os dados mostram isso acontecendo. As Marcas A e B passaram de 99% antes de N = 100. A Marca C oscilou entre 82% e 92% nas primeiras 50 respostas e ficou em torno de 91% a partir de N = 300. As marcas da faixa média oscilaram até N = 200 e depois se acomodaram. Entre a janela de N = 100 a 200 e a de N = 900 a 1.000, o desvio-padrão da frequência acumulada caiu de 60% a 95% para as dez marcas principais.

4.2 Experimento 2 — o que muda num mercado fragmentado

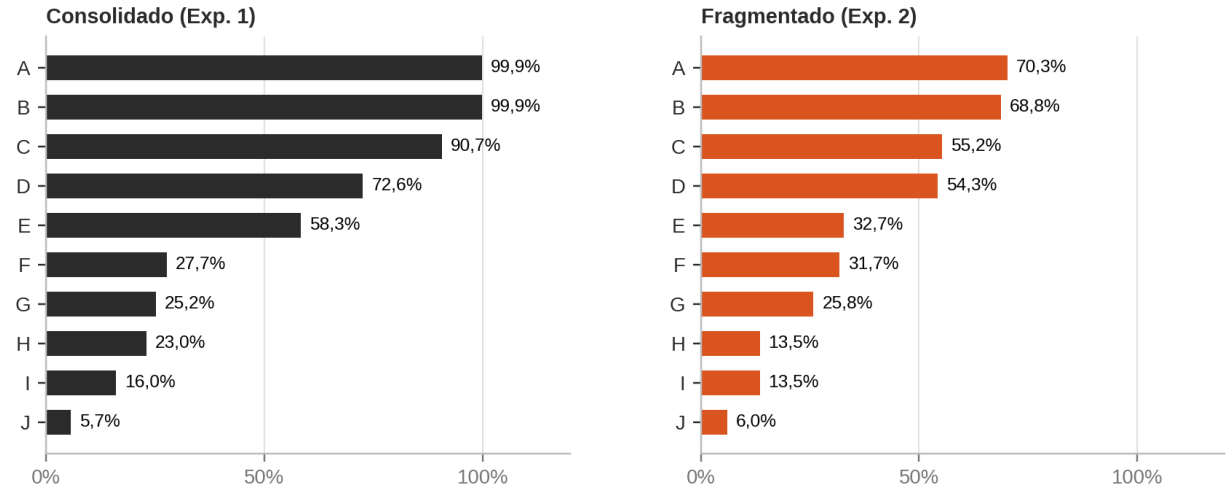
O segundo experimento repetiu o método numa categoria de perfil oposto: dezenas de marcas legítimas, nacionais e internacionais, competindo sem dominantes absolutas. Foram 1.000 chamadas, que citaram 144 marcas distintas, quase quatro vezes o Experimento 1.

Achado 2.1 — A forma da distribuição muda; a espinha se repete

No mercado fragmentado nenhuma marca se aproxima de 99,9%: a líder aparece em 70,3% das respostas e a segunda em 68,8%. A queda também é outra. No consolidado há um degrau entre o núcleo e a faixa média (a quinta marca aparece em 58%, a sexta em 28%). No fragmentado há uma rampa: as quatro primeiras vão de 70% a 54%, e as três seguintes ficam entre 25% e 33%.

No mercado fragmentado, nenhuma marca passa de 71%

Presença das 10 marcas mais frequentes em cada experimento (1.000 chamadas cada). Letras independentes por experimento.



Fonte: FA-MET-01 v1.0 · Exp. 1 e Exp. 2

Figura 3 — Presença das 10 marcas mais frequentes em cada experimento.

Tabela 2 — Comparação estrutural entre os Experimentos 1 e 2

Métrica	Consolidado	Fragmentado
Marcas distintas em 1.000 chamadas	38	144
Presença da 1ª marca	99,9%	70,3%
Presença da 2ª marca	99,9%	68,8%
Presença da 3ª marca	90,7%	55,2%
Presença da 5ª marca	58,3%	32,7%
Marcas com presença ≥ 50%	5	4
Marcas com presença ≥ 25%	7	7
Marcas com presença ≥ 10%	9	9
Marcas vistas só 1 ou 2 vezes	14	119
Top 5 — % do volume de menções	78%	70%
Top 10 — % do volume de menções	96%	93%

Base: 1.000 respostas por experimento.

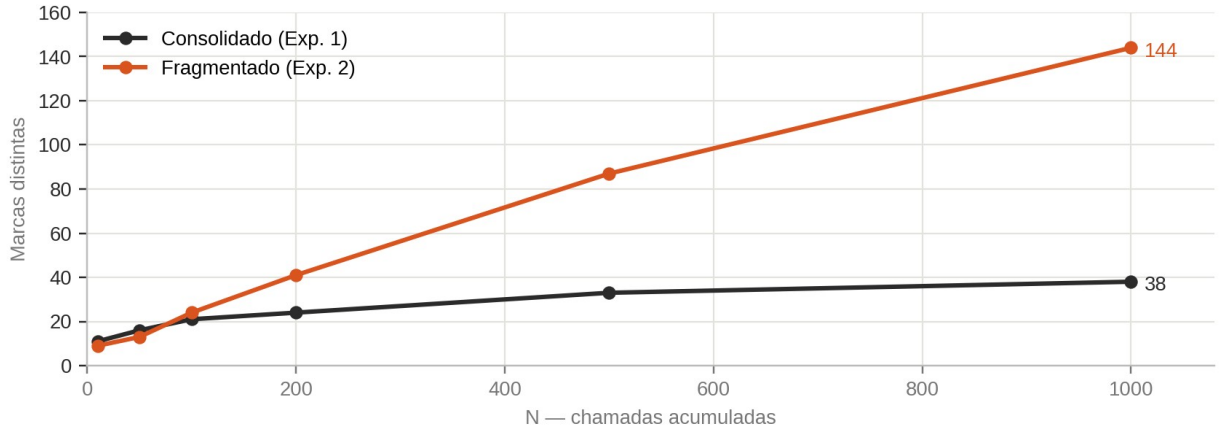
O mercado fragmentado tem 3,8 vezes mais marcas, mas o número de marcas acima de 25% (sete) e acima de 10% (nove) é idêntico nos dois. Existe uma espinha estrutural que se repete: a hierarquia está mais distribuída no fragmentado, em faixas de presença mais baixas, com a mesma quantidade de marcas em cada degrau.

Achado 2.2 — A cauda longa só é longa no mercado fragmentado

Em N = 100 os dois experimentos tinham visto um número parecido de marcas (21 e 24). A partir de N = 200 eles se separam: o consolidado tem 24 marcas, perto da saturação, e o fragmentado passa de 40 e segue crescendo até 144 em N = 1.000.

A cauda só é longa no mercado fragmentado

Marcas distintas observadas conforme N cresce, nos dois experimentos. Pontos medidos.



Fonte: FA-MET-01 v1.0 · Exp. 1 e Exp. 2

Figura 4 — Marcas distintas observadas conforme N cresce, nos dois experimentos.

Tabela 3 — Marcas distintas observadas por N

N	Consolidado	Fragmentado
10	11	9
50	16	13
100	21	24
200	24	41
500	33	87
1.000	38	144

Base: 1.000 respostas por experimento, em ordem de coleta.

Há ao menos dois regimes de vocabulário. No consolidado, ele é fechado, com saturação visível antes de N = 500. No fragmentado, é amplo e continua crescendo além de 1.000. O Experimento 3 testa até onde.

Achado 2.3 — O volume capturado pelo topo é parecido nos dois mercados

Apesar da diferença no topo, as cinco marcas mais frequentes concentram 78% das menções no consolidado e 70% no fragmentado; as dez mais frequentes, 96% e 93% (Tabela 2). Para quem quer entender quem domina e mapear a concorrência, as dez primeiras marcas capturam quase o mesmo volume nos dois tipos de mercado. A diferença real está na cauda.

4.3 Experimento 3 — uma série de 5.000 chamadas

O Experimento 2 deixou uma pergunta aberta: em 1.000 chamadas, a cauda do mercado fragmentado ainda crescia. O universo é aberto ou só grande? Para responder, fizemos uma série de 5.000 chamadas idênticas, com a mesma pergunta e o mesmo método do Experimento 2.

> *A hierarquia medida em 1.000 chamadas se mantém em 5.000? E o universo de marcas do mercado fragmentado é aberto ou apenas grande?*

Achado 3.1 — A hierarquia de 1.000 chamadas ficou dentro da margem

Comparamos a presença das dez marcas principais em N = 1.000 e em N = 5.000. Todas as variações ficaram abaixo de 3 pontos; a maior foi a da Marca D, de 54,3% para 57,2%. Nenhuma marca saiu do intervalo de 95% calculado em N = 1.000.

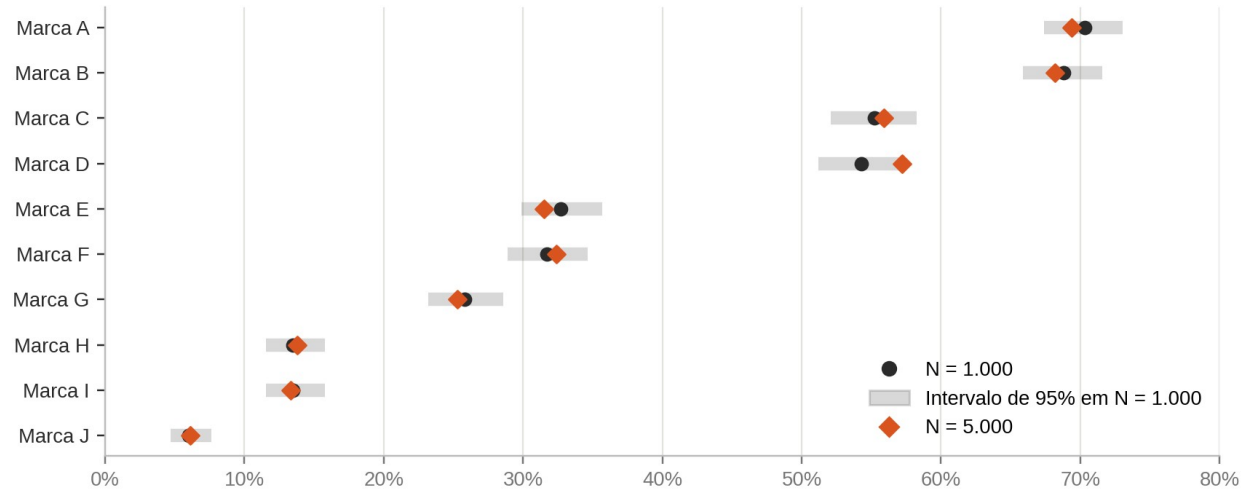
Tabela 4 — Presença das 10 marcas principais em N = 1.000 e N = 5.000

Marca	N = 1.000	IC 95% em 1.000	N = 5.000	Δ (pp)	Dentro do IC?
Marca A	70,3%	67,4% a 73,1%	69,4%	−0,9	Sim
Marca B	68,8%	65,9% a 71,6%	68,2%	−0,6	Sim
Marca C	55,2%	52,1% a 58,3%	55,9%	+0,7	Sim
Marca D	54,3%	51,2% a 57,4%	57,2%	+2,9	Sim
Marca E	32,7%	29,9% a 35,7%	31,5%	−1,2	Sim
Marca F	31,7%	28,9% a 34,6%	32,4%	+0,7	Sim
Marca G	25,8%	23,2% a 28,6%	25,3%	−0,5	Sim
Marca H	13,5%	11,5% a 15,8%	13,8%	+0,3	Sim
Marca I	13,5%	11,5% a 15,8%	13,3%	−0,2	Sim
Marca J	6,0%	4,7% a 7,6%	6,1%	+0,1	Sim

Base: 1.000 e 5.000 respostas. Letras pela ordem de presença em N = 1.000 da série, fixas depois de atribuídas.

Quatro mil chamadas a mais não tiraram nenhuma marca da margem de N = 1.000

Presença das 10 marcas mais frequentes do mercado fragmentado em N = 1.000 e N = 5.000.



Fonte: FA-MET-01 v1.0 · Exp. 3, série de 5.000 chamadas

Figura 5 — Presença em N = 1.000 (com intervalo de 95%) e em N = 5.000.

É exatamente o que a teoria prevê. Para uma marca com 55% de presença, a margem de 95% em 1.000 chamadas é de ±3,1 pontos; a Marca D variou 2,9. Quatro mil chamadas a mais refinaram as estimativas, mas não moveram a hierarquia. A hipótese do estudo se sustenta: a presença se comporta como proporção binomial, com a precisão da teoria clássica.

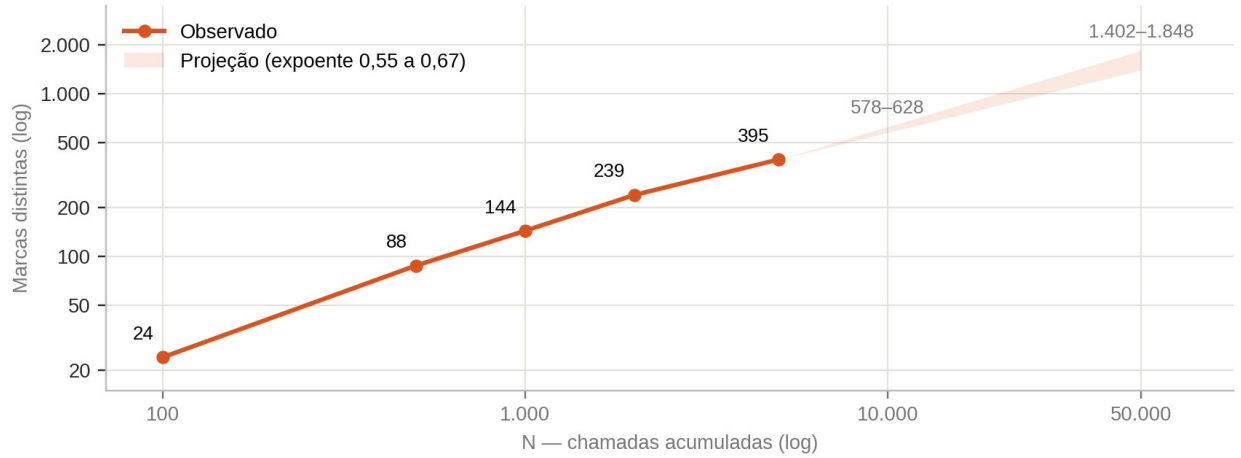
Este achado não permite afirmar que variações de poucos pontos entre duas rodadas de 1.000 chamadas sejam ruído: comparar duas rodadas soma a incerteza das duas (ver Seção 4.4).

Achado 3.2 — A cauda cresce, com retornos decrescentes

A hierarquia estava completa em 1.000 chamadas; o universo, não. Na série de 5.000 chamadas, o número de marcas distintas passou de 144 em N = 1.000 para 395 em N = 5.000, 2,7 vezes mais com 5 vezes o esforço.

A cauda do mercado fragmentado ainda cresce em 5.000, cada vez mais devagar

Marcas distintas observadas na série de 5.000 chamadas e faixa de projeção por lei de potência.



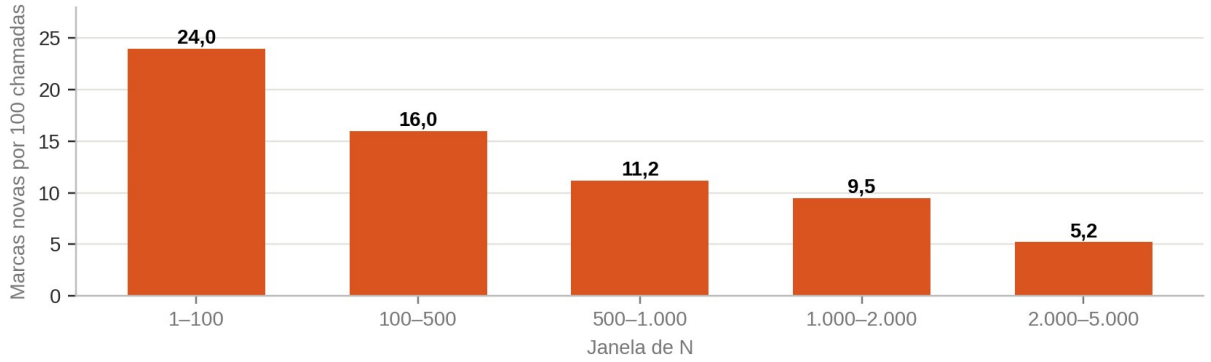
Fonte: FA-MET-01 v1.0 · Exp. 3. Expoente local caiu de 0,73 (1.000→2.000) para 0,55 (2.000→5.000).

Figura 6 — Marcas distintas na série de 5.000 chamadas e faixa de projeção, em escala logarítmica.

O ritmo de descoberta cai. Nas primeiras 100 chamadas surgiram 24 marcas novas; entre 2.000 e 5.000, cerca de 5 a cada 100 chamadas.

A cada 100 chamadas, surgem cada vez menos marcas novas

Marcas novas por 100 chamadas em cada janela da série de 5.000.



Fonte: FA-MET-01 v1.0 · Exp. 3, série de 5.000 chamadas

Figura 7 — Marcas novas por 100 chamadas em cada janela da série de 5.000.

O crescimento do vocabulário segue aproximadamente uma lei de potência, a lei de Heaps (Heaps, 1978), conhecida da recuperação de informação, mas o expoente local diminui: foi de 0,73 entre 1.000 e 2.000 chamadas e de 0,55 entre 2.000 e 5.000. Projetando com expoentes nessa faixa, 10.000 chamadas revelariam entre 580 e 630 marcas, e 50.000, entre 1.400 e 1.850. O universo é grande, a taxa de descoberta nunca zera, mas cada nova faixa de chamadas rende menos.

As projeções são extrapolações de cinco pontos medidos e servem para dimensionar coleta, não para afirmar o tamanho do vocabulário do modelo.

Achado 3.3 — Parte da cauda de amostra pequena não se confirma

Nas primeiras 1.000 chamadas da série, 119 marcas apareceram só uma ou duas vezes. As 4.000 chamadas seguintes permitem testar se elas fazem parte do vocabulário do modelo: uma marca real com presença baixa teria quatro vezes mais oportunidades de reaparecer.

71% das marcas raras de N = 1.000 voltaram nas 4.000 chamadas seguintes

119 marcas vistas só 1 ou 2 vezes nas primeiras 1.000 chamadas da série de 5.000.



Fonte: FA-MET-01 v1.0 · Exp. 3, série de 5.000 chamadas

Figura 8 — Marcas raras das primeiras 1.000 chamadas que reapareceram nas 4.000 seguintes.

Das 119, 85 (71%) reapareceram. As outras 34 (29%) não voltaram. Parte delas pode ser captura pontual, como modelo de produto sem marca-mãe ou entidade mal extraída; parte pode ser marca real com presença tão baixa (perto de 0,02%) que 4.000 chamadas não bastam para revê-la. Os dados não separam os dois casos.

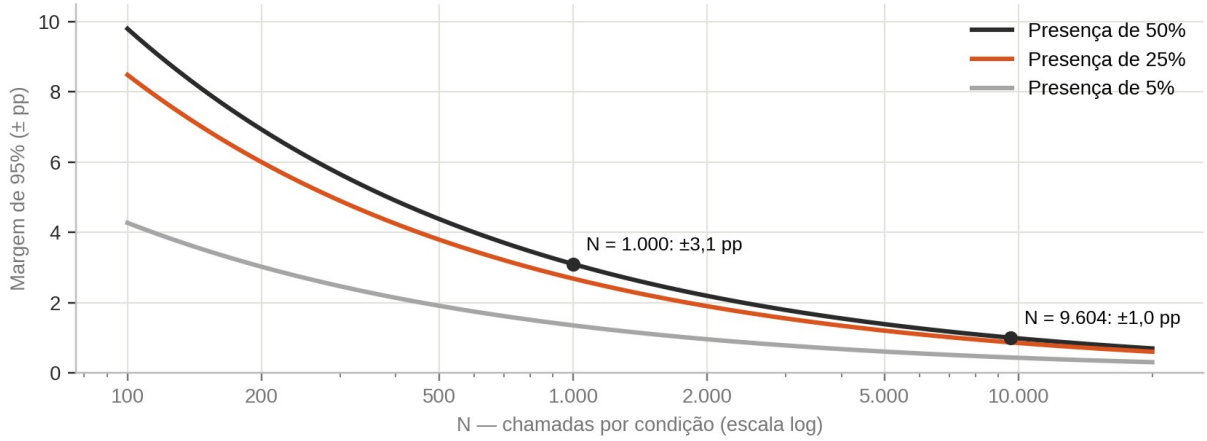
A implicação prática não depende dessa separação: uma lista de cauda tirada de 1.000 chamadas tem cerca de um terço de marcas não confirmadas e precisa vir com esse aviso.

4.4 Quantas chamadas são necessárias

A precisão de uma medida de presença depende só de N e da própria presença. A Figura 9 mostra a margem de 95% conforme N cresce.

Com 1.000 chamadas, a margem é de ±3 pp; para ±1 pp, são precisas ~9.600

Meia-largura do intervalo de 95% para a presença de uma marca, por N e por nível de presença.



FA-MET-01 v1.0 · cálculo binomial: 1,96 × √(p(1−p)/N)

Figura 9 — Margem de 95% da presença de uma marca, por N e por nível de presença.

Tabela 5 — N recomendado por objetivo

Objetivo	N por condição	O que entrega
Diagnóstico inicial	100 a 200	Regime (concentrado ou fragmentado) e topo aproximado, com ±7 a 10 pp

Objetivo	N por condição	O que entrega
Hierarquia visível	1.000	Presença com até ±3 pp; ordem confiável entre marcas separadas por mais de ~6 pp
Mudança entre rodadas	1.000 por rodada	Mudanças acima de ~4,4 pp (presença perto de 50%) são distinguíveis de ruído
Precisão de ±1 pp	~9.600	Presença com ±1 pp em qualquer nível
Cauda até 0,1%	~3.000	95% de chance de ver ao menos uma vez cada marca com presença de 0,1%
Cauda até 0,02%	~15.000	Idem para 0,02%

Margens para 95% de confiança. Linhas de cauda: $1 - (1 - p)^N \geq 0{,}95$.

Um raciocínio comum leva a esperar precisão maior que essa: como as respostas de uma mesma categoria se parecem, cada chamada valeria por várias. Não é o caso. Chamadas em conversas novas são independentes, e a semelhança entre respostas já está contida na própria presença. Também não serve de evidência o N a partir do qual a média acumulada fica a ±1 ponto do valor final da mesma série: essa métrica converge por construção, porque as últimas chamadas fazem parte do valor final. O Experimento 3 confirma a teoria clássica: 1.000 chamadas dão ±3 pontos, não ±1.

5. O que se generaliza e o que varia

Com lastro nos três experimentos

- **Convergência por amostragem repetida.** A presença de cada marca se estabiliza conforme N cresce, como prevê a lei dos grandes números, em qualquer categoria.
- **Hierarquia mensurável e estável.** Algumas marcas aparecem sistematicamente mais que outras, e multiplicar a amostra por cinco manteve a ordem dentro da margem.
- **Precisão binomial.** A margem de 95% segue a fórmula clássica: ±3,1 pontos em 1.000 chamadas, ±1 ponto perto de 9.600.
- **Cauda de amostra pequena pouco confiável.** Marcas vistas uma ou duas vezes em 1.000 chamadas incluem uma parcela relevante que não se confirma com mais volume.

Varia por categoria e exige medição própria

- **Forma do topo.** Duas marcas em 99,9% no consolidado; nenhuma acima de 71% no fragmentado.
- **Tamanho do universo.** 38 contra 144 marcas em 1.000 chamadas.
- **Saturação.** O consolidado satura antes de 1.000 chamadas; o fragmentado ainda cresce em 5.000.
- **Composição por ambiente.** As marcas abaixo do núcleo podem mudar conforme o ambiente de acesso ao modelo.

6. Implicações por leitor

Para marcas e profissionais de marketing

Estar no vocabulário ativo do modelo deixou de ser detalhe técnico. Marcas do núcleo têm presença contínua sem custo de mídia; marcas fora dele enfrentam invisibilidade estrutural. A primeira pergunta de um monitoramento é se a categoria tem núcleo e quem está nele, e 100 a 200 chamadas respondem isso. A segunda é a posição da marca, e exige rodadas de 1.000 chamadas, lembrando que só mudanças acima de cerca de 4 pontos entre rodadas são distinguíveis de ruído.

Para pesquisadores e jornalistas

Uma única pergunta a uma IA é uma única jogada de moeda. A marca citada pode ser uma escolha que aparece em 95% das respostas ou um sorteio que aparece em 1%. Qualquer afirmação sobre o que uma IA recomenda precisa de centenas de amostras independentes, idealmente 1.000 ou mais, repetidas periodicamente. Atribuir viés com base em marcas da cauda rara pede cuidado: parte dela não se confirma com mais volume.

Para quem desenvolve produtos com IA

Se um produto depende de uma IA recomendar algo de forma consistente, a consistência precisa ser medida, não assumida. Algumas escolhas são quase determinísticas, acima de 90%; outras variam em 30% ou mais das respostas. Perguntas mais específicas, instruções de formato e restrições de domínio reduzem a variação, mas o primeiro passo é medir o nível atual com o protocolo deste estudo.

7. Limitações

- **Um ambiente.** Os três experimentos foram coletados na interface do ChatGPT. A regularidade estatística deve valer para outros modelos e ambientes de acesso; a composição das marcas não necessariamente.
- **Modelo na interface.** A interface não permite fixar nem registrar o modelo que responde. Não é possível garantir que todas as chamadas tenham sido atendidas pelo mesmo modelo.
- **Duas categorias.** Os números específicos (núcleo de cinco marcas, 29% de cauda não confirmada, expoente da curva de descoberta) refletem estas categorias.
- **Validade temporal.** Atualizações do modelo podem mudar o ranking. As conclusões sobre a forma das curvas e a precisão tendem a ser mais estáveis que os números por marca.
- **Uma pergunta por experimento.** Outras formulações da mesma intenção podem gerar distribuições diferentes. Esse efeito não foi medido aqui.
- **Uma data.** Todas as coletas são de 27 de abril de 2026.
- **Transparência.** O modelo, os textos das perguntas e as categorias não são divulgados, e a base não é pública. Os números podem ser conferidos entre si, mas não refeitos por terceiros com os mesmos insumos.
- **Normalização.** A filtragem de termos não-marca afeta os números absolutos, sobretudo da cauda. A lista é conservadora: na dúvida, o termo foi mantido.

8. Reprodutibilidade

O que se transfere é o protocolo: número de chamadas, conversa nova por chamada, pergunta fixa, presença binária por resposta, normalização por marca-mãe e intervalos binomiais. Com ele, qualquer equipe pode repetir o desenho num modelo e numa categoria de sua escolha e comparar a forma das curvas. Replicações em outros modelos, outras categorias, outros momentos e via API são as que mais refinam o que se generaliza.

9. Referências

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*, jan. 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*. Nova York: Academic Press.

TechCrunch (2025). ChatGPT users send 2.5 billion prompts a day. 21 jul. 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

Anexo A — Ficha técnica

Campo	Registro
Código e versão	FA-MET-01 · v1.0
Pergunta	Uma por experimento, fixa, em forma de intenção de consumo, sem marcas como âncora. Textos não divulgados
Categorias	Exp. 1: mercado consolidado. Exp. 2 e 3: mercado fragmentado (mesma categoria). Não divulgadas
Idioma	Português do Brasil
Ambiente	Interface do ChatGPT no navegador, por automação (web scraping), com busca na web habilitada
Data de coleta	27/04/2026
Isolamento	Conversa nova por chamada, sem histórico
N	Exp. 1: 1.000. Exp. 2: 1.000. Exp. 3: 5.000. Total: 7.000 chamadas
Unidade de análise	Resposta × marca presente (presença binária por resposta)
Normalização	Consolidação por marca-mãe; varejistas, marketplaces e termos genéricos filtrados; na dúvida, o termo foi mantido
Anonimização	Letras pela ordem de presença em N = 1.000 de cada experimento, fixas depois de atribuídas
Estatística	Intervalos de 95% por Wilson; limite de análise individual em 0,3% de presença

Anexo B — Histórico de versões

Data	Versão	O que mudou
30/04/2026	1.0	Primeira versão publicada.