ESTUDO METODOLÓGICO · FA-MET-02

API ou interface, eis a questão

Por que a mesma pergunta traz marcas diferentes conforme o ambiente de acesso

---

Base: 1.000 chamadas via API e 1.000 via interface do ChatGPT · mesma pergunta · mesma janela de coleta (27/04/2026)

Versão 1.0 · 30 de abril de 2026

Rodrigo Beckmann
Cofundador e CTO
First Answer

1. Resumo executivo

A mesma pergunta, com a mesma ferramenta de busca, traz marcas diferentes conforme chega pela API ou pela interface do ChatGPT. Medimos 1.000 chamadas em cada ambiente, em paralelo, e encontramos cinco diferenças.

1. **Só duas marcas não mudam.** As Marcas B e C aparecem em quase 100% das respostas nos dois ambientes. A Marca A, também em 100% na API, cai para 77,5% no navegador. Da quarta posição em diante, as marcas trocam de lugar: a Marca D vai de 11,1% na API a 92,6% no navegador; a Marca F, de 66,5% a 6,4%.
2. **O universo mal se sobrepõe.** A API citou 39 marcas e o navegador 37, mas só 20 aparecem nos dois. Das 56 marcas observadas, 36 (64%) são exclusivas de um ambiente.
3. **A busca explica parte da diferença.** O navegador buscou na web em 99,5% das chamadas; a API, em 77,5%. Dentro da própria API, algumas marcas só aparecem quando há busca e outras só quando não há.
4. **Forçar a busca na API não resolveria.** Comparando só as chamadas com busca nos dois ambientes, a Marca D continua em 93,1% no navegador e 0,3% na API.
5. **O jeito de buscar também difere.** Quando busca, a API lê em média 22,6 URLs e faz 1,55 consultas por resposta; o navegador, 16,9 URLs em uma consulta.

Das 15 marcas mais relevantes, 12 têm diferença entre ambientes cujo intervalo de 95% não inclui zero. Não é ruído de amostragem: com 1.000 chamadas por ambiente, a margem de cada presença é de no máximo ±3,1 pontos percentuais, e as diferenças chegam a 82.

> **O que isso muda**
>
> A API é porta de integração para desenvolvedores, não o lugar onde o consumidor encontra a IA. Quem mede presença de marca via API e apresenta o resultado como o que o usuário vê está medindo outro sistema. Todo monitoramento precisa declarar o ambiente medido, e a troca de um pelo outro não se resolve com ajuste de configuração.

2. A pergunta

O usuário real não usa API. Ele abre o ChatGPT no navegador ou no aplicativo, digita a pergunta e lê a resposta. Se o objetivo é saber o que o consumidor vê, a medição precisa usar o ambiente dele ou provar que outro ambiente entrega resposta equivalente.

A intuição comum diz que a equivalência existe: mesmo modelo, mesma pergunta, mesma distribuição de probabilidade, e a interface seria só uma camada visual sobre o mesmo backend.

Não encontramos estudo publicado que teste essa equivalência. Fishkin e O'Donnell (2026) mostraram que as recomendações de marca variam a cada execução e também entre plataformas e entre recursos de uma mesma plataforma, mas compararam produtos diferentes (ChatGPT, Claude, IA do Google), com perguntas feitas manualmente por voluntários, e deixaram em aberto se chamadas via API produzem a mesma variação. A documentação da OpenAI informa que, na API, o modelo decide sozinho se usa a ferramenta de busca, a menos que o desenvolvedor a force (OpenAI, 2026). Esse comportamento, se diferente do da interface, já bastaria para os ambientes divergirem.

> *A interface do ChatGPT e a API, com busca habilitada, produzem amostras estatisticamente equivalentes de marcas recomendadas?*

Hipótese testada: os dois ambientes são equivalentes.

Critério de refutação: a hipótese cai se alguma marca com presença acima de 10% em um dos ambientes tiver diferença entre ambientes cujo intervalo de 95% não inclua zero. Com 1.000 chamadas por ambiente, isso equivale a diferenças acima de cerca de 4 pontos percentuais.

3. Método

Foram 1.000 chamadas em cada ambiente, feitas em paralelo em 27 de abril de 2026, numa janela de poucas horas, para que atualizações de índice ou de roteamento não afetassem um lado só. Cada chamada abriu uma conversa nova, sem histórico. Na API, a busca na web estava habilitada, com o modelo livre para decidir quando usá-la, temperatura padrão e sem instrução de sistema customizada. Na interface, as chamadas foram feitas sem login, com geolocalização no Brasil. A coleta foi desenhada para usar o mesmo modelo nos dois ambientes; como a interface sem login não permite fixar nem registrar o modelo, isso não pode ser verificado chamada a chamada (ver Limitações). A pergunta, uma intenção de consumo em português do Brasil escrita como um usuário digitaria, foi idêntica nas 2.000 chamadas.

De cada resposta extraímos todas as marcas citadas, por um processo misto de regras e modelo de linguagem, sem distinguir menção principal de secundária. Variações da mesma marca-mãe foram consolidadas e varejistas, bandeiras genéricas e termos não-marca foram filtrados. Para cada chamada também registramos se o modelo buscou na web, quantas URLs recuperou e quantas consultas fez.

Por que 1.000 chamadas. Chamadas em conversas novas são amostras independentes, então a presença de cada marca segue a distribuição binomial. Com 1.000 chamadas, a margem de 95% é de no máximo ±3,1 pontos percentuais (para presença de 50%) e menor nos extremos. É precisão suficiente para este estudo, cujas diferenças relevantes passam de 10 pontos. Os intervalos usam o método de Wilson (Wilson, 1927) para cada presença e o de Newcombe (Newcombe, 1998) para a diferença entre ambientes. O estudo FA-MET-01 detalha a escolha de N por objetivo (Beckmann, 2026).

Anonimização. As marcas são identificadas por letras atribuídas pela ordem de presença na API. A Marca A da API é a mesma Marca A do navegador. O foco é a estrutura da divergência, não o ranking da categoria. A ficha técnica completa está no Anexo A.

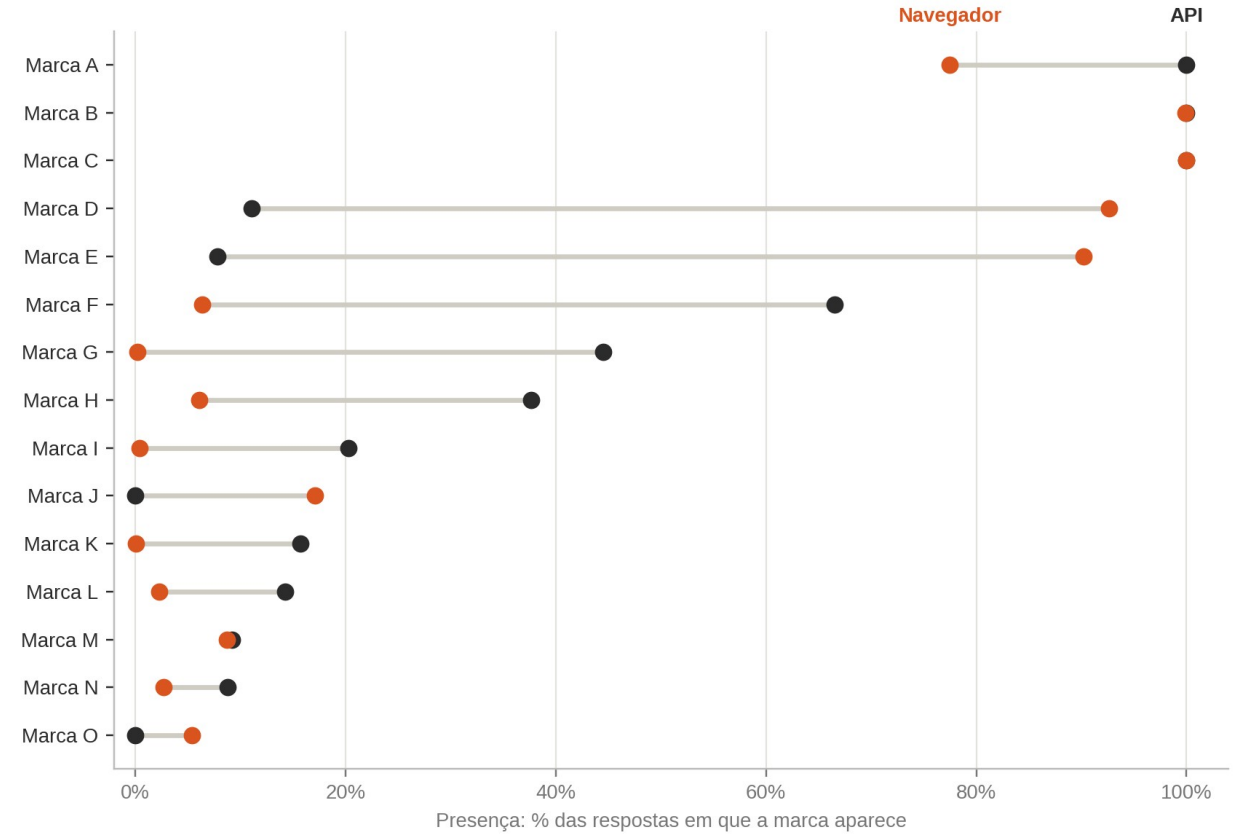
4. Achados

Achado 1 — Duas marcas sobrevivem à troca de ambiente; o resto não

As três marcas mais frequentes na API são as mesmas do topo do navegador, mas em graus diferentes. As Marcas B e C ficam perto de 100% nos dois ambientes: é o núcleo determinístico da categoria, que nenhuma variação de ambiente tira da resposta. A Marca A atinge 100% na API e cai para 77,5% no navegador, uma diferença de −22,5 pontos.

Fora de B e C, a presença muda de ambiente para ambiente

Presença das 15 marcas mais relevantes na API (preto) e no navegador (laranja). Margem de 95%: até ±3,1 pp por ponto.



Fonte: FA-MET-02 v1.0 · 1.000 chamadas por ambiente · mesmo modelo, mesma pergunta, mesma janela

Figura 1 — Presença das 15 marcas mais relevantes em cada ambiente. A linha liga a mesma marca nos dois ambientes; quanto mais longa, maior a divergência.

Tabela 1 — Presença das 15 marcas mais relevantes, por ambiente

Marca	API	Navegador	Δ (pp)	IC 95% da Δ (pp)
Marca A	100,0%	77,5%	−22,5	−25,2 a −20,0
Marca B	100,0%	99,9%	−0,1	−0,6 a +0,3
Marca C	100,0%	100,0%	0,0	−0,4 a +0,4
Marca D	11,1%	92,6%	+81,5	+78,7 a +83,8
Marca E	7,8%	90,2%	+82,4	+79,7 a +84,7
Marca F	66,5%	6,4%	−60,1	−63,3 a −56,7
Marca G	44,5%	0,2%	−44,3	−47,4 a −41,2
Marca H	37,7%	6,1%	−31,6	−34,9 a −28,2
Marca I	20,3%	0,4%	−19,9	−22,5 a −17,4

Marca	API	Navegador	Δ (pp)	IC 95% da Δ (pp)
Marca J	0,0%	17,1%	+17,1	+14,9 a +19,6
Marca K	15,7%	0,1%	−15,6	−18,0 a −13,4
Marca L	14,3%	2,3%	−12,0	−14,4 a −9,7
Marca M	9,2%	8,7%	−0,5	−3,0 a +2,0
Marca N	8,8%	2,7%	−6,1	−8,2 a −4,1
Marca O	0,0%	5,4%	+5,4	+4,1 a +7,0

Base: 1.000 respostas por ambiente. Δ = navegador − API. Intervalo da diferença pelo método de Newcombe (Wilson). Só B, C e M têm intervalo que inclui zero.

Da quarta posição em diante, as marcas que ocupam o ranking no navegador não são as da API. A Marca D aparece em 92,6% das respostas do navegador e em 11,1% das da API. A Marca E, em 90,2% contra 7,8%. No sentido oposto, a Marca F aparece em 66,5% na API e 6,4% no navegador; a Marca G, em 44,5% contra 0,2%.

Não são diferenças de calibração, são inversões: o que um ambiente trata como recomendação principal, o outro trata como cauda quase invisível.

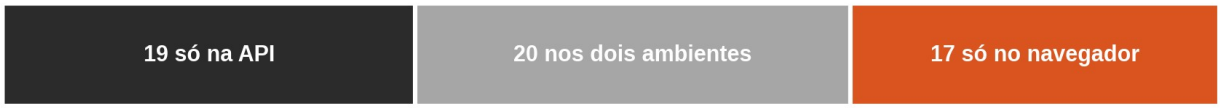
Este achado não permite afirmar que o navegador favorece alguma marca em geral: vale para esta categoria, este modelo e esta data.

Achado 2 — O universo de marcas tem tamanho parecido, mas mal se sobrepõe

Em 1.000 chamadas, a API citou 39 marcas distintas e o navegador, 37. Os totais parecidos sugerem vocabulários equivalentes, mas só 20 marcas aparecem nos dois conjuntos. Dezenove nunca surgiram no navegador e dezessete nunca surgiram na API.

Das 56 marcas observadas, só 20 aparecem nos dois ambientes (36%)

Universo de marcas distintas em 1.000 chamadas por ambiente: 39 na API, 37 no navegador.



Fonte: FA-MET-02 v1.0 · 1.000 chamadas por ambiente · mesmo modelo, mesma pergunta, mesma janela

Figura 2 — Universo de marcas observado em cada ambiente.

Nenhum ambiente conhece um superconjunto do outro, nem os dois conhecem o mesmo conjunto com pesos diferentes. Cada um tem vocabulário próprio, com uma região comum (o núcleo e parte da faixa média) e uma região exclusiva relevante. Coletar 1.000 chamadas num ambiente entrega o universo daquele ambiente, não o do modelo.

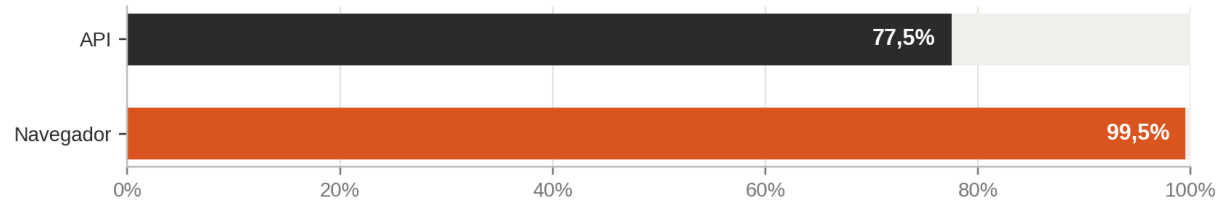
Este achado não permite estimar o universo total do modelo: marcas raras podem existir nos dois ambientes e não ter aparecido em 1.000 chamadas.

Achado 3 — A busca na web explica parte da divergência

Registramos, para cada chamada, se o modelo ativou a busca. No navegador, buscou em 99,5% das chamadas. Na API, com a mesma ferramenta habilitada, buscou em 77,5%: em 225 das 1.000 chamadas respondeu só com o conhecimento do treinamento.

A API respondeu sem buscar na web em 22,5% das chamadas

% das chamadas em que o modelo ativou a busca na web, por ambiente (busca habilitada nos dois).



Fonte: FA-MET-02 v1.0 · 1.000 chamadas por ambiente · mesmo modelo, mesma pergunta, mesma janela

Figura 3 — Frequência de busca na web em cada ambiente.

Dentro da própria API, as respostas com e sem busca citam marcas diferentes:

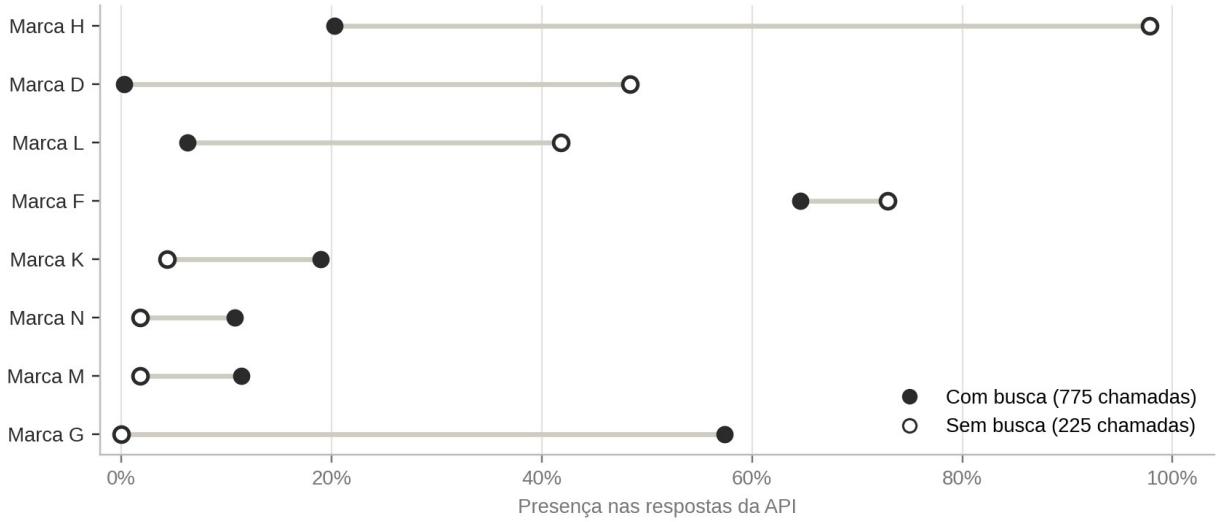
Tabela 2 — Presença na API, com e sem busca na web

Marca	Com busca	Sem busca	Δ (pp)
Marca H	20,3%	97,8%	+77,5
Marca D	0,3%	48,4%	+48,1
Marca L	6,3%	41,8%	+35,5
Marca F	64,6%	72,9%	+8,3
Marca K	19,0%	4,4%	−14,6
Marca N	10,8%	1,8%	−9,0
Marca M	11,4%	1,8%	−9,6
Marca G	57,4%	0,0%	−57,4

Base: 775 respostas com busca e 225 sem busca. Δ = sem busca − com busca. Margem de 95%: até ±3,5 pp (com busca) e ±6,5 pp (sem busca).

Dentro da própria API, buscar ou não buscar troca as marcas

Presença na API conforme a chamada ativou ou não a busca, para as marcas com diferença relevante.



Fonte: FA-MET-02 v1.0 · margem de 95%: até ±3,5 pp (com busca) e ±6,5 pp (sem busca)

Figura 4 — Presença na API conforme a chamada ativou ou não a busca.

A Marca H aparece em 97,8% das chamadas sem busca e em 20,3% das com busca. A Marca D vai de 0,3% com busca a 48,4% sem. No sentido oposto, a Marca G só aparece quando há busca (57,4%

contra 0%). Cada fonte tem seu vocabulário: com busca, a resposta segue o da web; sem busca, o do treinamento. Como a API busca em três quartos das chamadas e o navegador em quase todas, parte da divergência vem dessa mistura de regimes.

Não buscar não é defeito: na API, a decisão de usar a busca é do próprio modelo, conforme a documentação da OpenAI (OpenAI, 2026). O que o experimento mostra é que essa decisão não é estável. Para a mesma pergunta, em 1.000 chamadas, o modelo busca em algumas e não em outras, e o resultado agregado combina os dois regimes.

Achado 4 — Forçar a busca em 100% das chamadas da API não resolveria

Se a divergência viesse só da frequência de busca, bastaria forçar a API a buscar sempre, o que a própria API permite (OpenAI, 2026). O experimento permite aproximar esse teste sem nova coleta: comparamos só as 775 respostas da API com busca e as 995 do navegador com busca.

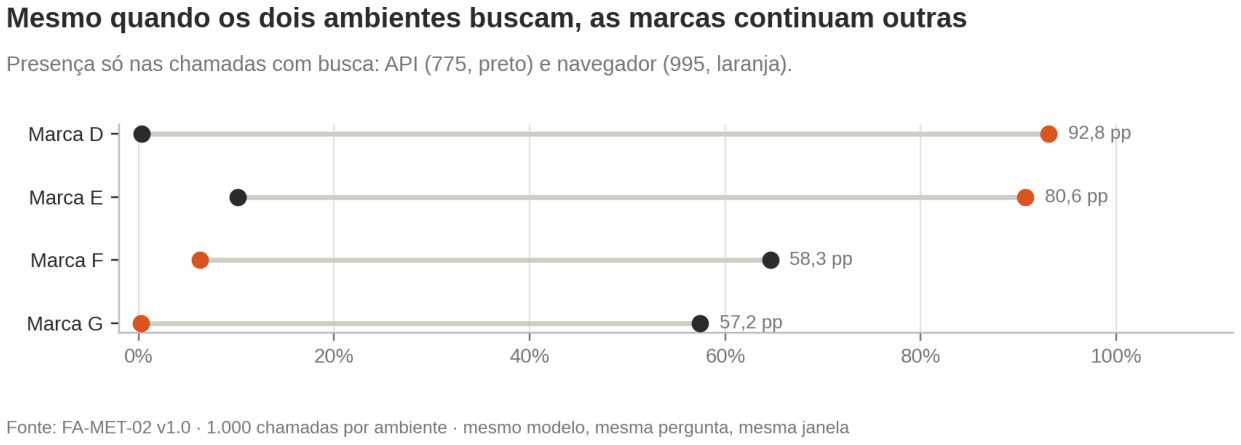


Figura 5 — Presença nas chamadas com busca nos dois ambientes, para as marcas de maior divergência.

A divergência continua. A Marca D aparece em 93,1% das respostas do navegador com busca e em 0,3% das da API com busca, uma diferença ainda maior que no total. A Marca E fica em 90,7% contra 10,1%. A Marca F, em 6,3% no navegador contra 64,6% na API; a Marca G, em 0,2% contra 57,4%. Nesse recorte, a API citou 29 marcas e o navegador 34, com só 16 em comum.

Forçar a busca reduziria a divergência, porque eliminaria a parcela que vem do treinamento, mas não a eliminaria. Duas hipóteses, possivelmente combinadas, explicam o resíduo: sistemas de busca diferentes (índices, motores ou pesos de fonte distintos) e instruções de sistema diferentes em cada ambiente. O experimento não testa nenhuma das duas; as duas são compatíveis com o padrão observado.

Achado 5 — Quando os dois buscam, buscam de jeitos diferentes

Tabela 3 — Comportamento de busca nas chamadas com busca

Métrica	API	Navegador
Chamadas com busca	775 (77,5%)	995 (99,5%)
URLs recuperadas por resposta — média	22,6	16,9
URLs recuperadas por resposta — mediana	26	18
Consultas de busca por resposta — média	1,55	~1

Base: chamadas com busca em cada ambiente.

A API explora mais: recupera mais URLs e, em cerca de metade das chamadas, faz mais de uma consulta antes de responder. O navegador faz praticamente sempre uma única consulta. A diferença não é só buscar ou não buscar, é também a profundidade da busca, e ela ajuda a explicar o resíduo do Achado 4.

5. O que se generaliza e o que varia

Deve valer em outras categorias e modelos

- **Ambientes acessam vocabulários parcialmente disjuntos.** Os mecanismos identificados (decisão autônoma de buscar, infraestrutura de busca e instruções distintas) são características gerais de modelos com ferramentas.
- **O núcleo determinístico é o que mais resiste.** Marcas que dominam a categoria tendem a aparecer em qualquer ambiente.
- **A decisão de buscar não é estável no agregado.** Para a mesma pergunta, o modelo busca em parte das chamadas e não em outras.

Específico desta coleta

- As magnitudes: 22,5% de chamadas sem busca, divergências de até 82 pontos no total e 93 no recorte com busca.
- Quais marcas mudam e em que direção.
- O tamanho do universo e da sobreposição (56 marcas, 20 em comum).

6. Implicações por leitor

Para quem mede presença de marca via API

A coleta via API é, no melhor caso, um proxy do ambiente do usuário, e este experimento mostra que esse proxy pode estar longe do original. Em um quarto das chamadas o modelo nem busca e responde do treinamento, que tem vocabulário próprio. O retrato que sai da API mistura dois regimes e não corresponde a nenhum deles isoladamente.

Há uso legítimo para a API: medir o que o modelo sabe do treinamento, sem a web do momento. Serve para estudar se uma marca está consolidada na memória do modelo ou para auditar atualizações. O que não se pode é apresentar esse retrato como o que o usuário vê.

Para quem mede via interface

A coleta pela interface é a forma adequada de medir o que o consumidor vê, porque é o ambiente dele. É mais cara em escala, mais sujeita a limites de uso e mais difícil de automatizar; são problemas de engenharia, não de método. Quem mede pela interface precisa registrar que mede o regime em que quase sempre há busca, e que os achados podem não valer para integrações sem busca.

Para quem desenvolve produtos com IA

Trocar de interface para API, ou o contrário, é mudança de produto, não de operação: as marcas que o sistema menciona mudam. Equipes que prototipam pela interface e vão para produção pela API podem ver o comportamento mudar, e isso não é defeito de implementação.

Para a série de estudos da First Answer

O estudo FA-MET-01 foi coletado pela interface do ChatGPT, o mesmo ambiente do navegador deste estudo. Suas conclusões estruturais (convergência por amostragem, hierarquia mensurável, núcleo determinístico em mercado consolidado) e a composição das marcas refletem o que o consumidor vê. Replicado via API, o estudo deve mostrar curvas da mesma forma, com outras marcas abaixo do núcleo.

7. Limitações

- **Uma categoria.** Mercado consolidado, com poucos players relevantes. Em mercados fragmentados a magnitude pode ser maior, menor ou ter outra forma.
- **Um modelo.** Outros modelos com busca devem apresentar o mesmo tipo de divergência, mas a magnitude precisa ser medida modelo a modelo.
- **Um momento.** Mudanças na forma como a API ou a interface acionam a busca ou priorizam fontes podem mudar os números. O achado que deve durar é a existência da divergência e suas origens estruturais, não a sua magnitude.
- **Uma pergunta.** Uma única formulação de intenção de consumo. Outras formulações podem gerar outros padrões.
- **Modelo na interface.** Sem login, a interface do ChatGPT não permite escolher nem registrar o modelo que respondeu. Não é possível descartar que parte da divergência venha do roteamento de modelos da interface, e não só do ambiente.
- **Uma data.** Todas as chamadas são de 27 de abril de 2026.
- **Transparência.** O modelo, o texto da pergunta e a categoria não são divulgados, e a base não é pública.
- **Recorte com busca parcial.** A comparação restrita às chamadas com busca foi reportada para as marcas de maior divergência; as demais marcas não foram tabuladas nesse recorte.

8. Reprodutibilidade

O que se transfere é o protocolo: 1.000 chamadas via API com busca habilitada e 1.000 via interface, mesma pergunta, mesma janela de tempo, conversas novas, registro de busca por chamada e intervalos binomiais para cada presença e cada diferença. Com ele, qualquer equipe pode repetir o desenho no modelo e na categoria de sua escolha. Replicações em mercados fragmentados, em outros modelos e com a interface logada (que permite registrar o modelo) são as mais úteis para refinar o achado.

9. Referências

Beckmann, R. (2026). Como medir o que uma IA recomenda (FA-MET-01, v1.0). First Answer.

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/

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8), 873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

OpenAI (2026). Web search. *OpenAI API Documentation*. https://developers.openai.com/docs/guides/tools-web-search.

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-02 · v1.0
Pergunta	Intenção de consumo em categoria consolidada de produto, texto idêntico nas 2.000 chamadas. Texto e categoria não divulgados
Idioma	Português do Brasil
API	Busca na web habilitada, com o modelo decidindo quando usá-la; temperatura padrão; sem instrução de sistema
Interface	ChatGPT sem login, geolocalizado no Brasil; modelo não fixável nem registrável
Busca web	Registrada por chamada: ativou ou não, nº de URLs, nº de consultas
Data de coleta	27/04/2026, em paralelo nos dois ambientes
Isolamento	Conversa nova por chamada, sem histórico
N	1.000 chamadas por ambiente
Unidade de análise	Resposta × marca presente (presença binária por resposta)
Extração	Mista (regras e modelo de linguagem); todas as marcas citadas, sem distinção de menção principal ou secundária
Normalização	Consolidação por marca-mãe; varejistas, bandeiras genéricas e termos não-marca filtrados
Anonimização	Letras pela ordem de presença na API, iguais nos dois ambientes
Estatística	Intervalos de 95% por Wilson; diferenças por Newcombe

Anexo B — Histórico de versões

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