METHODOLOGICAL STUDY · FA-MET-02

# API or Interface: That Is the Question

Why the same question returns different brands depending on how it is asked

---

**Base:** 1,000 runs via API and 1,000 via the ChatGPT interface · same question · same collection window (April 27, 2026)

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

**Rodrigo Beckmann**
Co-founder and CTO
First Answer

# 1. Executive summary

The same question, with the same search tool, returns different brands depending on whether it arrives through the API or the ChatGPT interface. We ran 1,000 queries in each environment, in parallel, and found five differences.

1. **Only two brands stay put.** Brands B and C appear in nearly 100% of responses in both environments. Brand A, also at 100% in the API, drops to 77.5% in the browser. From the fourth position on, brands trade places: Brand D goes from 11.1% in the API to 92.6% in the browser; Brand F, from 66.5% to 6.4%.
2. **The universes barely overlap.** The API mentioned 39 brands and the browser 37, but only 20 appear in both. Of the 56 brands observed, 36 (64%) are exclusive to one environment.
3. **Search explains part of the difference.** The browser searched the web in 99.5% of runs; the API, in 77.5%. Within the API itself, some brands appear only when it searches and others only when it does not.
4. **Forcing search in the API would not fix it.** Comparing only runs with search in both environments, Brand D still stands at 93.1% in the browser and 0.3% in the API.
5. **The way they search differs too.** When it searches, the API reads 22.6 URLs on average and issues 1.55 queries per response; the browser, 16.9 URLs in a single query.

Of the 15 most relevant brands, 12 show a difference between environments whose 95% interval excludes zero. This is not sampling noise: with 1,000 runs per environment, the margin on each presence is at most ±3.1 percentage points, and the differences reach 82.

> **Why it matters**
>
> The API is an integration gateway for developers, not where consumers meet the AI. Anyone who measures brand presence via the API and presents the result as what users see is measuring a different system. Every monitoring program must state the environment it measures, and swapping one for the other is not fixed by a configuration tweak.

## 2. The question

Real users do not use the API. They open ChatGPT in a browser or the app, type the question and read the answer. If the goal is to know what consumers see, measurement has to use their environment, or prove that another environment delivers an equivalent answer.

Common intuition says the equivalence holds: same model, same question, same probability distribution, with the interface being just a visual layer on top of the same backend.

We found no published study testing this equivalence. Fishkin and O'Donnell (2026) showed that brand recommendations vary from run to run, and also across platforms and across features of the same platform, but they compared different products (ChatGPT, Claude, Google's AI), with questions asked manually by volunteers, and left open whether API calls produce the same variation. OpenAI's documentation states that, in the API, the model decides on its own whether to use the search tool, unless the developer forces it (OpenAI, 2026). If that behavior differs from the interface's, it alone would be enough for the environments to diverge.

> *Do the ChatGPT interface and the API, with search enabled, produce statistically equivalent samples of recommended brands?*

**Hypothesis tested:** the two environments are equivalent.

**Refutation criterion:** the hypothesis fails if any brand with presence above 10% in either environment shows a between-environment difference whose 95% interval excludes zero. With 1,000 runs per environment, that corresponds to differences above roughly 4 percentage points.

## 3. Method

We ran 1,000 queries in each environment, in parallel on April 27, 2026, within a window of a few hours, so that index or routing updates could not affect only one side. Each run opened a new conversation with no history. In the API, web search was enabled, with the model free to decide when to use it, default temperature and no custom system instruction. In the interface, runs were made logged out, geolocated in Brazil. Collection was designed to use the same model in both environments; since the logged-out interface does not allow the model to be fixed or recorded, this cannot be verified run by run (see Limitations). The question, a purchase intent in Brazilian Portuguese written as a user would type it, was identical across all 2,000 runs.

From each response we extracted every brand mentioned, through a mixed process of rules and a language model, with no distinction between primary and secondary mentions. Variants of the same parent brand were consolidated, and retailers, generic banners and non-brand terms were filtered out. For each run we also recorded whether the model searched the web, how many URLs it retrieved and how many queries it issued.

**Why 1,000 runs.** Runs in new conversations are independent samples, so each brand's presence follows the binomial distribution. With 1,000 runs, the 95% margin is at most ±3.1 percentage points (at 50% presence) and smaller at the extremes. That is enough precision for this study, whose relevant differences exceed 10 points. Intervals use the Wilson method (Wilson, 1927) for each presence and the Newcombe method (Newcombe, 1998) for the difference between environments. Study FA-MET-01 details the choice of N by objective (Beckmann, 2026).

**Anonymization.** Brands are identified by letters assigned in order of presence in the API. Brand A in the API is the same Brand A in the browser. The focus is the structure of the divergence, not the category ranking. The full study sheet is in Appendix A.

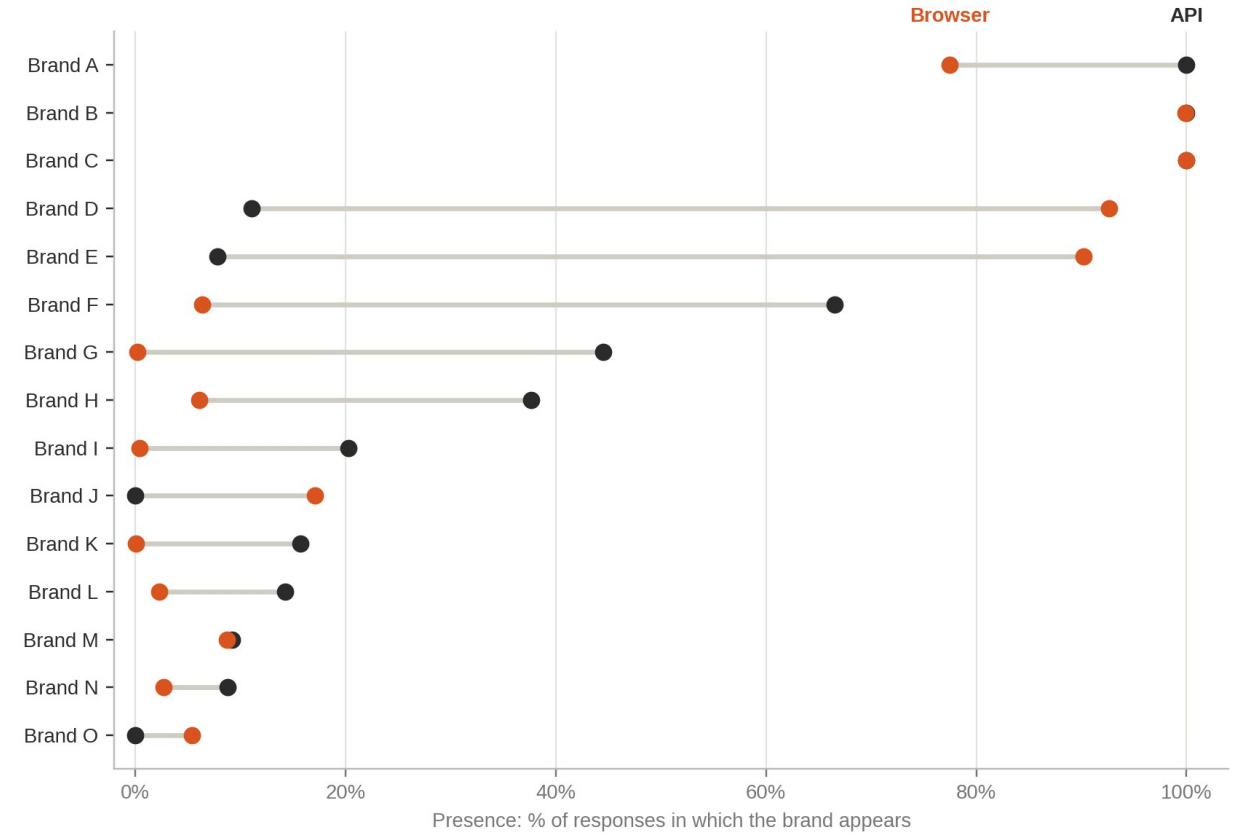
# 4. Findings

## Finding 1 — Two brands survive the change of environment; the rest do not

The three most frequent brands in the API are the same ones at the top in the browser, but to different degrees. Brands B and C stay close to 100% in both environments: this is the category's deterministic core, which no change of environment removes from the response. Brand A reaches 100% in the API and drops to 77.5% in the browser, a difference of −22.5 points.

### Outside B and C, presence changes from one environment to the other

Presence of the 15 most relevant brands in the API (black) and the browser (orange). 95% margin: up to ±3.1 pp per point.




Source: FA-MET-02 v1.0 · 1,000 runs per environment · same question, same window

*Figure 1 — Presence of the 15 most relevant brands in each environment. The line joins the same brand across the two environments; the longer it is, the larger the divergence.*

**Table 1 — Presence of the 15 most relevant brands, by environment**

| Brand | API | Browser | Δ (pp) | 95% CI of Δ (pp) |
|---|---|---|---|---|
| Brand A | 100.0% | 77.5% | −22.5 | −25.2 to −20.0 |
| Brand B | 100.0% | 99.9% | −0.1 | −0.6 to +0.3 |
| Brand C | 100.0% | 100.0% | 0.0 | −0.4 to +0.4 |
| Brand D | 11.1% | 92.6% | +81.5 | +78.7 to +83.8 |
| Brand E | 7.8% | 90.2% | +82.4 | +79.7 to +84.7 |
| Brand F | 66.5% | 6.4% | −60.1 | −63.3 to −56.7 |
| Brand G | 44.5% | 0.2% | −44.3 | −47.4 to −41.2 |
| Brand H | 37.7% | 6.1% | −31.6 | −34.9 to −28.2 |
| Brand I | 20.3% | 0.4% | −19.9 | −22.5 to −17.4 |

| Brand | API | Browser | Δ (pp) | 95% CI of Δ (pp) |
|---|---|---|---|---|
| Brand J | 0.0% | 17.1% | +17.1 | +14.9 to +19.6 |
| Brand K | 15.7% | 0.1% | −15.6 | −18.0 to −13.4 |
| Brand L | 14.3% | 2.3% | −12.0 | −14.4 to −9.7 |
| Brand M | 9.2% | 8.7% | −0.5 | −3.0 to +2.0 |
| Brand N | 8.8% | 2.7% | −6.1 | −8.2 to −4.1 |
| Brand O | 0.0% | 5.4% | +5.4 | +4.1 to +7.0 |

Base: 1,000 responses per environment. Δ = browser − API. Interval of the difference by the Newcombe (Wilson) method. Only B, C and M have intervals that include zero.

From the fourth position on, the brands that make up the browser ranking are not the API's. Brand D appears in 92.6% of browser responses and in 11.1% of API responses. Brand E, in 90.2% versus 7.8%. In the opposite direction, Brand F appears in 66.5% in the API and 6.4% in the browser; Brand G, in 44.5% versus 0.2%.

These are not calibration differences, they are inversions: what one environment treats as a lead recommendation, the other treats as an almost invisible tail.

This finding does not show that the browser favors any brand in general: it holds for this category, this model and this date.

## Finding 2 — The brand universes are similar in size but barely overlap

In 1,000 runs, the API mentioned 39 distinct brands and the browser 37. The similar totals suggest equivalent vocabularies, but only 20 brands appear in both sets. Nineteen never surfaced in the browser and seventeen never surfaced in the API.

**Of the 56 brands observed, only 20 appear in both environments (36%)**

Universe of distinct brands in 1,000 runs per environment: 39 in the API, 37 in the browser.




Source: FA-MET-02 v1.0 · 1,000 runs per environment · same question, same window

*Figure 2 — Brand universe observed in each environment.*

Neither environment knows a superset of the other, nor do both know the same set with different weights. Each has its own vocabulary, with a shared region (the core and part of the middle band) and a sizeable exclusive region. Collecting 1,000 runs in one environment yields that environment's universe, not the model's.

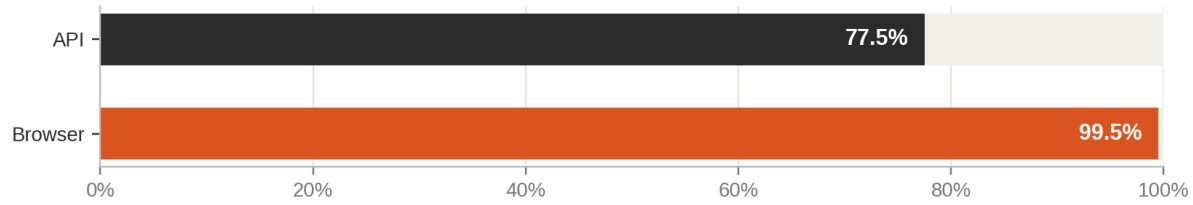
This finding does not allow estimating the model's total universe: rare brands may exist in both environments and not have appeared in 1,000 runs.

## Finding 3 — Web search explains part of the divergence

For each run we recorded whether the model triggered search. In the browser, it searched in 99.5% of runs. In the API, with the same tool enabled, it searched in 77.5%: in 225 of the 1,000 runs it answered from training knowledge alone.

### The API answered without searching the web in 22.5% of runs

% of runs in which the model triggered web search, by environment (search enabled in both).




Source: FA-MET-02 v1.0 · 1,000 runs per environment · same question, same window

*Figure 3 — Web search frequency in each environment.*

Within the API itself, responses with and without search mention different brands:

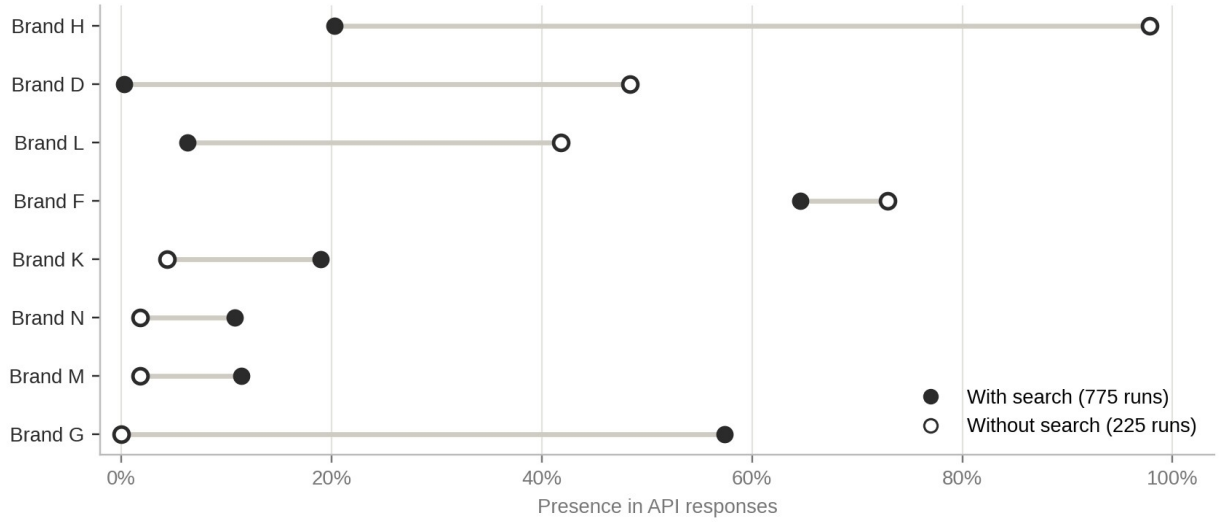
**Table 2 — Presence in the API, with and without web search**

| Brand | With search | Without search | Δ (pp) |
|---|---|---|---|
| Brand H | 20.3% | 97.8% | +77.5 |
| Brand D | 0.3% | 48.4% | +48.1 |
| Brand L | 6.3% | 41.8% | +35.5 |
| Brand F | 64.6% | 72.9% | +8.3 |
| Brand K | 19.0% | 4.4% | −14.6 |
| Brand N | 10.8% | 1.8% | −9.0 |
| Brand M | 11.4% | 1.8% | −9.6 |
| Brand G | 57.4% | 0.0% | −57.4 |

Base: 775 responses with search and 225 without. Δ = without search − with search. 95% margin: up to ±3.5 pp (with search) and ±6.5 pp (without).

### Within the API itself, searching or not changes the brands

Presence in the API depending on whether the run triggered search, for brands with a relevant difference.




Source: FA-MET-02 v1.0 · 95% margin: up to ±3.5 pp (with search) and ±6.5 pp (without)

*Figure 4 — Presence in the API depending on whether the run triggered search.*

Brand H appears in 97.8% of runs without search and in 20.3% of runs with search. Brand D goes from 0.3% with search to 48.4% without. In the opposite direction, Brand G appears only when there is

search (57.4% versus 0%). Each source has its own vocabulary: with search, the response follows the web's; without it, the training data's. Since the API searches in three quarters of runs and the browser in almost all, part of the divergence comes from this mix of regimes.

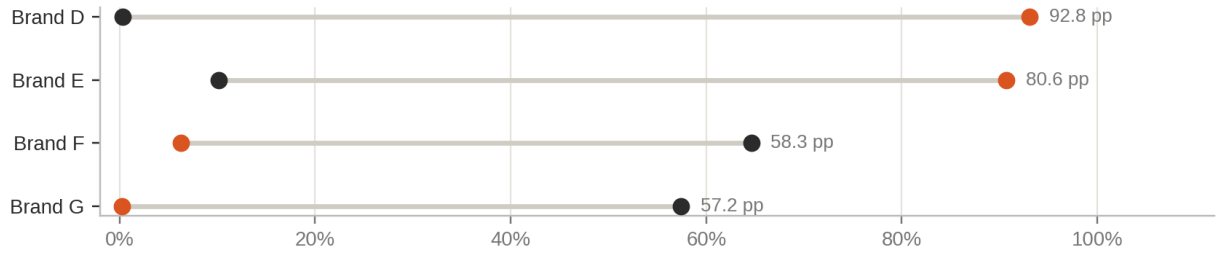
Not searching is not a defect: in the API, the decision to search belongs to the model itself, according to OpenAI's documentation (OpenAI, 2026). What the experiment shows is that this decision is not stable. For the same question, over 1,000 runs, the model searches in some and not in others, and the aggregate result blends the two regimes.

## Finding 4 — Forcing search in 100% of API runs would not fix it

If the divergence came only from search frequency, forcing the API to always search would be enough, and the API allows it (OpenAI, 2026). The experiment lets us approximate that test without new data: we compared only the 775 API responses with search and the 995 browser responses with search.

**Even when both environments search, the brands stay different**

Presence in runs with search only: API (775, black) and browser (995, orange).




Source: FA-MET-02 v1.0 · 1,000 runs per environment · same question, same window

*Figure 5 — Presence in runs with search in both environments, for the brands with the largest divergence.*

The divergence persists. Brand D appears in 93.1% of browser responses with search and in 0.3% of API responses with search, an even larger difference than overall. Brand E stands at 90.7% versus 10.1%. Brand F, at 6.3% in the browser versus 64.6% in the API; Brand G, at 0.2% versus 57.4%. In this subset, the API mentioned 29 brands and the browser 34, with only 16 in common.

Forcing search would reduce the divergence, because it would remove the share that comes from training, but it would not eliminate it. Two hypotheses, possibly combined, explain the remainder: different search systems (distinct indexes, engines or source weights) and different system instructions in each environment. The experiment tests neither; both are consistent with the observed pattern.

## Finding 5 — When both search, they search differently

**Table 3 — Search behavior in runs with search**

| Metric | API | Browser |
|---|---|---|
| Runs with search | 775 (77.5%) | 995 (99.5%) |
| URLs retrieved per response — mean | 22.6 | 16.9 |
| URLs retrieved per response — median | 26 | 18 |
| Search queries per response — mean | 1.55 | ~1 |

Base: runs with search in each environment.

The API explores more: it retrieves more URLs and, in about half of its runs, issues more than one query before answering. The browser almost always issues a single query. The difference is not only whether to search, but also how deep, and it helps explain the remainder in Finding 4.

# 5. What generalizes and what varies

### Should hold in other categories and models

- **Environments draw on partly disjoint vocabularies.** The mechanisms identified (autonomous search decisions, search infrastructure and distinct instructions) are general features of tool-using models.
- **The deterministic core is the most resilient.** Brands that dominate the category tend to appear in any environment.
- **The decision to search is not stable in aggregate.** For the same question, the model searches in some runs and not in others.

### Specific to this collection

- The magnitudes: 22.5% of runs without search, divergences of up to 82 points overall and 93 in the search-only subset.
- Which brands change, and in which direction.
- The size of the universe and of the overlap (56 brands, 20 shared).

# 6. Implications by reader

### For those measuring brand presence via the API

API collection is, at best, a proxy for the user's environment, and this experiment shows that the proxy can be far from the original. In a quarter of runs the model does not even search and answers from training, which has its own vocabulary. The picture that comes out of the API blends two regimes and matches neither on its own.

There is a legitimate use for the API: measuring what the model knows from training, without the live web. It serves to study whether a brand is established in the model's memory or to audit updates. What it cannot do is stand in for what the user sees.

### For those measuring via the interface

Collecting through the interface is the right way to measure what consumers see, because it is their environment. It is more expensive at scale, more subject to usage limits and harder to automate; those are engineering problems, not methodological ones. Anyone measuring through the interface should record that they are measuring the regime in which search almost always happens, and that findings may not hold for integrations without search.

### For AI product builders

Switching from interface to API, or the reverse, is a product change, not an operational one: the brands the system mentions change. Teams that prototype in the interface and ship through the API may see behavior change, and that is not an implementation bug.

### For First Answer's study series

Study FA-MET-01 was collected through the ChatGPT interface, the same environment as this study's browser side. Its structural conclusions (convergence through sampling, a measurable hierarchy, a deterministic core in a concentrated market) and its brand composition reflect what consumers see. Replicated via the API, it should show curves of the same shape, with other brands below the core.

# 7. Limitations

- **One category.** A concentrated market with few relevant players. In fragmented markets the magnitude may be larger, smaller or shaped differently.
- **One model.** Other models with search should show the same kind of divergence, but the magnitude must be measured model by model.
- **One moment.** Changes in how the API or the interface trigger search or prioritize sources may change the figures. What should last is the existence of the divergence and its structural origins, not its magnitude.
- **One question.** A single phrasing of a purchase intent. Other phrasings may produce other patterns.
- **Model in the interface.** Logged out, the ChatGPT interface does not allow choosing or recording the model that answered. We cannot rule out that part of the divergence comes from the interface's model routing, and not only from the environment.
- **One date.** All runs took place on April 27, 2026.
- **Transparency.** The model, the question wording and the category are not disclosed, and the dataset is not public.
- **Partial search-only subset.** The comparison restricted to runs with search was reported for the brands with the largest divergence; the remaining brands were not tabulated in that subset.

# 8. Reproducibility

What transfers is the protocol: 1,000 API runs with search enabled and 1,000 through the interface, same question, same time window, new conversations, search recorded per run, and binomial intervals for each presence and each difference. With it, any team can repeat the design on the model and category of its choice. Replications in fragmented markets, on other models and with a logged-in interface (which allows recording the model) are the most useful to refine the finding.

# 9. References


Beckmann, R. (2026). How to measure what an AI recommends (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*, 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/

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

# Appendix A — Study sheet

| Field | Record |
|---|---|
| Code and version | FA-MET-02 · v1.0 |
| Question | Purchase intent in a concentrated product category, identical wording across all 2,000 runs. Wording and category not disclosed |
| Language | Brazilian Portuguese |
| API | Web search enabled, with the model deciding when to use it; default temperature; no system instruction |
| Interface | ChatGPT logged out, geolocated in Brazil; model cannot be fixed or recorded |
| Web search | Recorded per run: triggered or not, number of URLs, number of queries |
| Collection date | April 27, 2026, in parallel in both environments |
| Isolation | New conversation per run, no history |
| N | 1,000 runs per environment |
| Unit of analysis | Response × brand present (binary presence per response) |
| Extraction | Mixed (rules and a language model); every brand mentioned, with no distinction between primary and secondary mentions |
| Normalization | Parent-brand consolidation; retailers, generic banners and non-brand terms filtered out |
| Anonymization | Letters in order of presence in the API, identical in both environments |
| Statistics | 95% Wilson intervals; differences by Newcombe |

# 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. |