How stable are AI brand recommendations? A reproducibility and fidelity study

GenSig Research #1 · July 2026 · gensig.app · Case study: investing platforms, Argentina · Dataset and code published with this article

---

TL;DR. We asked the same purchase-intent questions 2,500 times across five API configurations, and 30 times by hand in the consumer apps (ChatGPT web and Gemini web), for one category: investing platforms in Argentina. Three findings:

- **The bare LLM API doesn't measure what users see — it measures the model's memory.** It missed 3 of the 5 platforms real browser sessions consistently recommend, and inflated two that browsers almost never mention.
- **Reproducibility and fidelity are independent axes.** Temperature 0 made answers highly repeatable — of the *wrong* brand set. Determinism is not accuracy.
- **Search-augmented APIs mirror their own consumer surface.** OpenAI's search variant reproduced ChatGPT web's top-5 exactly; Gemini with grounding matched Gemini web best and recovered every core brand the bare API missed.

If you measure “AI visibility” for a brand — as a growing category of tools does — these results say the *how* changes the answer more than most dashboards admit.

Why this matters

More buyers start product research inside AI assistants, and a category of tools has emerged to measure whether an assistant recommends *your* brand. These tools split into two camps: those that query model **APIs**, and those that scrape the **consumer chat apps**. The scraping camp argues APIs don't reflect what users actually see. The API camp argues scraping is fragile and unreproducible. Both camps make empirical claims. We couldn't find published data testing them — so we ran the experiment.

Method

Category & prompts. One purchase-intent category: investing platforms in Argentina (a live local market with fast-moving entrants). Five frozen Spanish prompts paraphrasing the same intent (best platforms / app for beginners / stocks & CEDEARs / online broker / top platforms). Prompts contain no brand names.

API conditions — 100 runs per prompt per condition (500 per condition, N=2,500):

Condition	What it is
`bare`	gpt-4o-mini, default temperature — the naive measurement
`bare_t0`	same, temperature 0
`persona`	+ system prompt “you're answering a person who lives in Argentina”
`search`	gpt-4o-mini-search-preview (built-in web search)
`grounded`	gemini-flash-latest with Google Search grounding

Consumer anchor. 30 manual runs in the real apps: 5 prompts × 3 passes on different days × 2 surfaces (ChatGPT web, Gemini web), from a regular consumer account on a residential Argentine

connection, one fresh chat per prompt, first answer only.

Measurement. Brand detection by frozen alias dictionary (25 brands). Metrics pre-specified before running: mention rate with 95% Wilson intervals; Fleiss' kappa across runs (runs as raters, brands as binary items); mean pairwise Jaccard of the top-3 set; tie-corrected Kendall's W on core-brand rankings; Spearman ρ between each condition's mention-rate vector and each anchor's. Total API cost: under $10. Every raw response archived as JSONL.

Results

1 · The consumer surfaces agree on a stable core

Across 15 browser runs per surface, ChatGPT web recommended the same five platforms with near-perfect consistency: InvertirOnline (15/15), Balanz (15/15), Cocos (14/15), PPI (11/15), Bull Market (11/15). Gemini web shares the leaders with a broader tail of fintech wallets. Whatever “the AI recommends” means to a user, in this category it is *not* noise.

2 · The bare API tells a different story than the browser

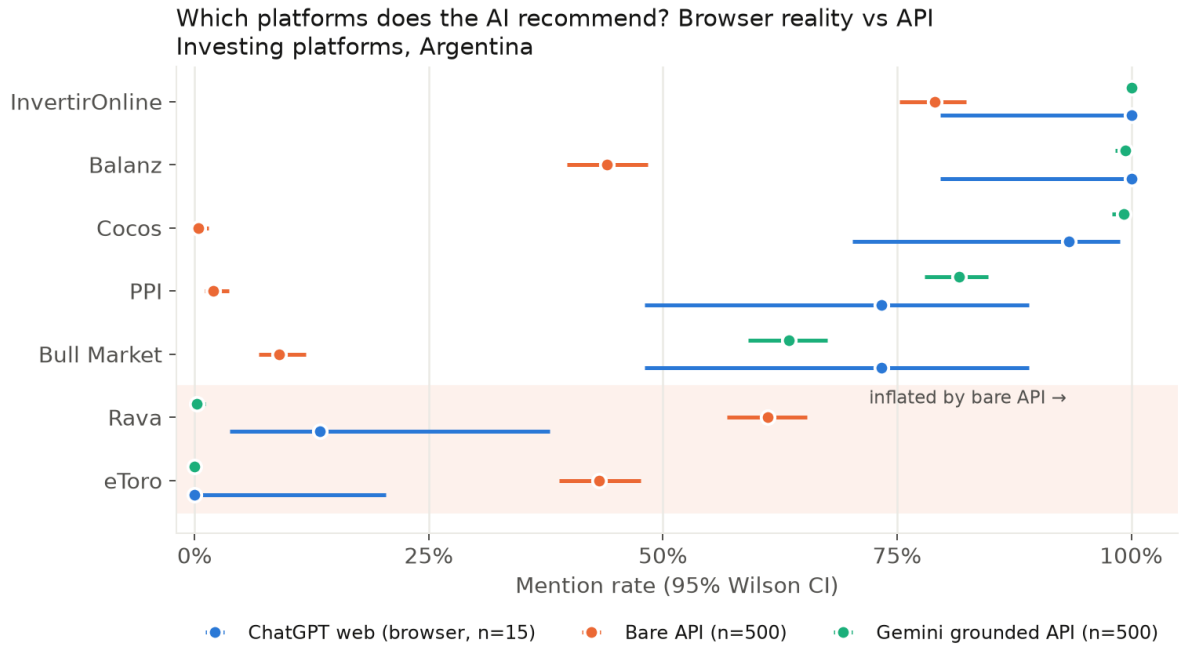

Figure 1 — Mention rates with 95% Wilson CIs: browser anchor vs bare API vs Gemini grounded.

The bare API (n=500) recovered InvertirOnline (79%) and Balanz (44%) — but **Cocos, the #3 browser recommendation, appears in under 1% of bare API runs.** PPI and Bull Market barely register. Meanwhile the bare API inflates **Rava (61%)** and **eToro (43%)**, which browsers mention in ≤2 of 15 runs. Rank correlation with the ChatGPT-web anchor: ρ = 0.23; top-5 overlap 25%.

Our interpretation: the bare model answers from parametric memory, which lags the market. Cocos' consumer rise is recent; the browser product, which searches the live web, reflects it — the model's weights don't. A geo persona prompt helped only marginally (ρ = 0.30).

3 · Determinism is not accuracy

At default temperature the bare API's top-3 set is volatile (mean pairwise Jaccard 0.36; modal top-3 in only 10% of runs). At temperature 0 it becomes highly repeatable (Jaccard 0.92, modal 84%) — **and its correlation with the browser anchor drops to ρ ≈ –0.04 / 0.10.** Temperature 0 gives you a precise, repeatable measurement of the model's outdated prior. “Consistent numbers” is often treated as

evidence of measurement quality; here, the most consistent configuration was the least faithful one.

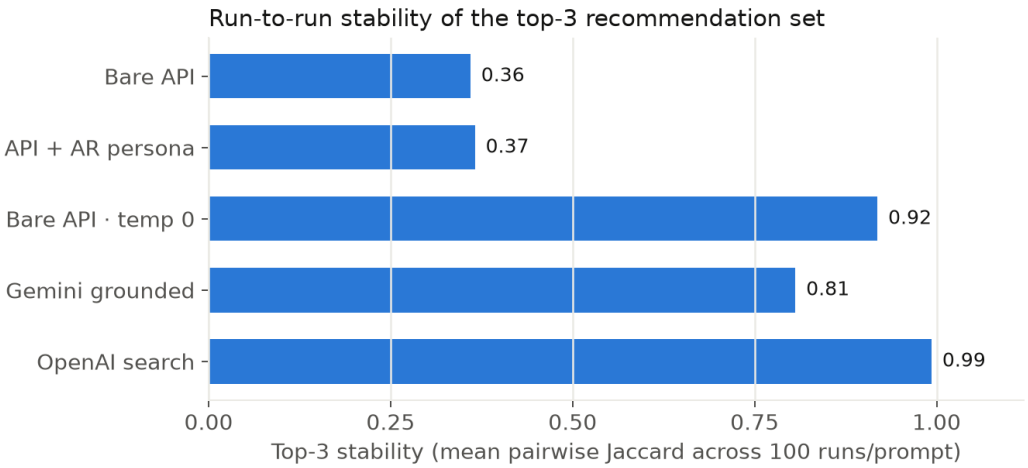

Figure 2 — Run-to-run stability of the top-3 set, by condition.

4 · Search-augmented APIs mirror their own consumer surface

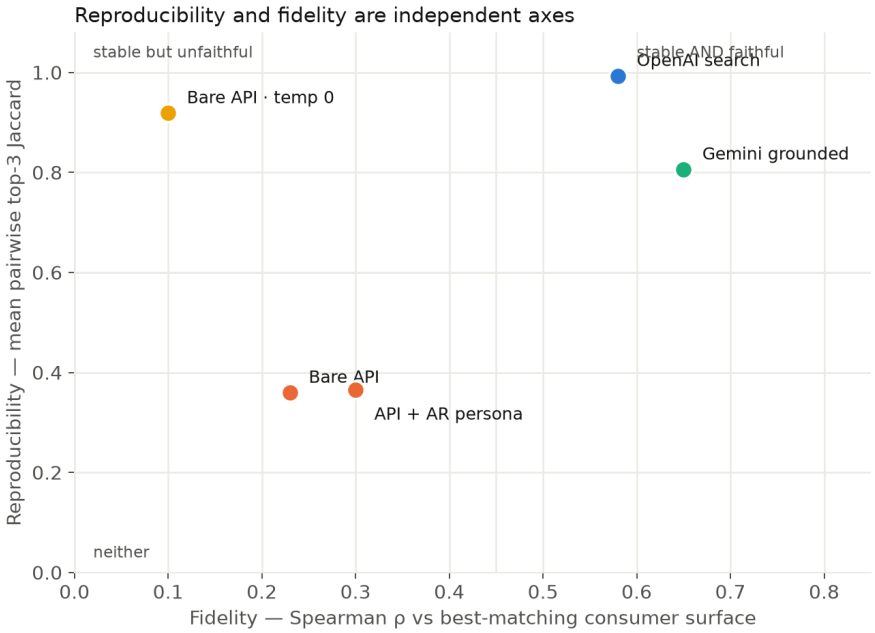

Figure 3 — Reproducibility × fidelity. Each search-augmented API lands in the useful quadrant.

- **OpenAI search-preview** reproduced the ChatGPT-web anchor's **top-5 exactly** (Jaccard 1.0; ρ = 0.58) — unsurprising, since ChatGPT web itself searched in most anchor runs.
- **Gemini grounding** was the best mirror of Gemini web (ρ = 0.65) and recovered **every** core brand the bare API missed (Cocos 99%, PPI 82%, Bull Market 63%).

Fidelity is **surface-specific**: an API-with-search approximates the consumer product of its own family. There is no single “what the AI says” — there is what ChatGPT says and what Gemini says, and each is best measured through its own search-augmented channel.

5 · Two cautionary observations

- **OpenAI search-preview's citations were reproducibly low-quality** in this category: the most-cited domains across 500 runs include an SEO content farm (79% of runs), an app-store mirror, and a German insurance forum. Its *brands* were faithful; its *sources* were not. It was also

near-frozen across runs (top-3 Jaccard 0.99), consistent with response caching.

- **Gemini grounding metadata was sparse:** only 3.4% of grounded runs exposed source chunks, so we treat Gemini source-level analysis as inconclusive in this study. Among the 17 runs that did expose sources, they were consistent and reasonable (rankia.com.ar 94%, reddit.com 65%).

Limitations

One category, one market, one time window (all API runs within hours; anchor across 3 days). The consumer anchor is 15 runs per surface from a single account and IP — enough to establish the stable core, not fine-grained rates. Brand detection is dictionary-based (aliases frozen pre-run). Model versions: gpt-4o-mini, gpt-4o-mini-search-preview, gemini-flash-latest (July 2026). Results may differ by category — a pilot on Argentine wallets showed smaller bare-API fidelity gaps; recency of a category's entrants appears to matter — by market, and as providers update models. Replications across categories are planned as Research #2.

Implications if you measure AI visibility

- **Don't measure with a bare API and call it reality.** In this category it misses recent entrants entirely and inflates stale ones.
- **Don't mistake repeatability for accuracy.** Temp-0 pipelines produce stable dashboards of the model's memory, not of what users see.
- **Measure per surface, through that surface's search-augmented channel.** OpenAI-with-search to approximate ChatGPT; Gemini-with-grounding to approximate Gemini.
- **Average over repeated runs, with intervals.** Single-shot measurements of a stochastic system are noise with a UI.

Data & code

The full raw dataset (2,530 responses, JSONL), the frozen prompt set, the brand dictionary, the runner and the analysis code are published alongside this article. Scraping-based tools cannot publish their equivalent — their raw data embeds ToS-violating collection. Reproducibility is the point.

GenSig (gensig.app) measures brand visibility in AI answers using repeated API sampling with confidence intervals — the methodology this study validates and stress-tests. Questions, replications and corrections welcome: support@gensig.app.