

The \$0 Split Test

How we ran a real 50/50 homepage A/B test with an AI agent, a cookie, and GA4 — no Optimizely, no VWO, no new contract. Here's the full recipe so you can reproduce it on your own stack.

2,366

USERS BUCKETED

14 days

FIXED WINDOW

2.3% → 4.4%

PER-USER CTR · P=0.006

2.7× → 1.8×

WEEK-1 LIFT → DURABLE

\$0

NEW SOFTWARE

INGREDIENTS

- ~60 lines of JS in a tiny plugin/snippet, injected early in <head>
- One first-party cookie to remember each visitor's variant (180 days, bots exempted)
- One GA4 custom dimension — user-scoped, carrying the variant label
- One GTM tag firing a click event for every CTA on both variants
- A fixed test window decided before launch — no peeking, no early calls
- An AI agent to write the bucketing code, register the dimensions, and run the math
- A human who owns the final call

\$0

total new tooling cost — the A/B SaaS category exists because instrumentation used to be expensive. It isn't anymore.

GOTCHAS WE HIT

- ! Instrument **surfaces, not labels**: our new page had 5 tracked CTAs, the old had 2 — asymmetry confounds raw click counts.
- ! Wire the cross-domain funnel **before** any conversion-based test — we measured intent, never signups.

THE RECIPE

1 Pick one primary metric + guardrails

Ours: CTA clicks (primary), with engagement rate, bounce, and session length as guardrails. The guardrails saved us from shipping a confident wrong answer.

2 Register a user-scoped variant dimension before launch

In GA4: Admin → Custom definitions → user scope. Without it you can never attribute per-user behavior to a variant. This is the step everyone skips.

3 Bucket 50/50 with a cookie

On first visit: `Math.random() < 0.5` → set cookie → variant B users get routed to the new page. Known bots exempted by user-agent. The cookie keeps every return visit consistent.

4 Instrument both variants equally

Same click event, same GTM tag, every CTA on both pages. Push the variant into the dataLayer on every event so the label rides along.

5 Commit to a fixed window, then wait

We ran 14 days. Week 1 showed a 2.7× lift; week 2 decayed to ~1.8×. The first week of any test flatters the new thing — novelty decay is real and free to measure.

6 Test significance on per-user rates

Two-proportion z-test: $\text{users-who-clicked} \div \text{users-bucketed}$, per variant. Per-user beats per-click (one enthusiastic clicker can't skew it). $p < 0.05$ on the primary, and read every guardrail too.

7 Make the tradeoff explicit — then let a human decide

Our new page won clicks but halved time-on-page. A test doesn't have to answer everything to be decision-grade; it has to make the tradeoff honest so a person can own the call. Keep rollback one flag-flip away.