CROspector: the deterministic engine we built, and the machine judges it was built for

This is Part 3 of a three-part story. Part 1 was the bet. Part 2 was the receipt. This is the answer.

The 3:47 a.m. audit

Every night, while you sleep, a crawler you did not invite opens your product pages (PDPs) and audits three prices for each one: **(1)** the price Googlebot reads on the page, **(2)** the price you sent to Merchant Center in your feed (which already carries your active `sale_price`), and **(3)** the price your checkout would actually charge if a shopper added the item to cart. You expect those three to agree, because you told Google about the promo through the feed and your checkout is consistent with it. Sometimes they do not — at 3:47 a.m., JavaScript injects your promotional banner a fraction of a second after Googlebot consumes the initial HTML, and the price the crawler records is the pre-promo price while your feed and your checkout already carry the discounted one. This is a failure mode Google itself documents: "Googlebot crawls the data present in the HTML returned from your web server" — and if your final price is rendered after that initial HTML, "this will trigger an error." By the time you wake up, every human-driven browser on the planet has rendered your page correctly, no shopper has encountered any broken state, and the partial render Googlebot consumed is not reproducible. But Google has the audit on your account; you have no copy of the bytes the crawler read and no way to reproduce them — and a machine has already decided what happens to that stock-keeping unit (SKU) next, on evidence you cannot inspect, while no human at your company is watching.

How often this is happening to your catalog right now is a question nobody can answer today. Not Google, which shows you the verdict but not the page state that produced it; not your team, which sees only the disapproval notice. Nothing in your stack records what the crawler saw at that moment. The mechanic Google calls "preemptive item disapproval" — removal on suspicion of mismatch, with a 28-day account-level review window during which the items remain suppressed — is the literal policy language, and its consequences are unambiguous: Google's own policy is that "disapproved products stop showing across Google." The affected SKU disappears from organic free listings (revenue you no longer earn) and from the paid Shopping and Performance Max inventories your media budget is bidding into (spend that is now either lost or rerouted at higher cost to a less-relevant SKU that does not convert at the same rate). The frequency is unknown; the asymmetry is not.

This is not an anecdote about Google being strict. It is the first appearance of a species: the **machine judge** — an algorithmic auditor that reads commerce surfaces per record, renders verdicts on a clock measured in machine time, not yours, and attaches the penalty directly to revenue. Google's shopping crawler is the incumbent judge. The buying agents now being wired

into checkout and payment protocols are the next bench: in October 2025, Visa announced its *Trusted Agent Protocol*, an ecosystem-led framework for AI agents to transact on merchant sites, with Adyen, Checkout.com, Cloudflare, Mastercard, Microsoft, Shopify, Stripe, and Worldpay among the named launch collaborators. Machines increasingly decide whether your products can be found, shown, and bought — and they read your store the way machines read: literally, repeatedly, unforgivingly.

Part 1 of this series was a bet. Part 2 was the receipt. This part is the answer: what we built, and the market the machine judges are creating for it.

What the receipt settled

Part 1 argued that conversion is not an optimization problem but a **feasibility** problem: a journey converts only if it is worth starting, possible to complete, and reliable under real-world constraints. Part 2 published the receipt of our first attempt to instrument that feasibility at scale: $32,341 of AWS spend running a state-of-the-art LLM agent across 10.5 million product pages on 1,056 French retailers. Completion rate: 9%. Not because retailer sites blocked us — we got past roughly 92% of the bot-protection systems guarding those retailers' sites: Cloudflare, Akamai, DataDome, and the rest of the "verify you are human" gauntlet. The 9% is what an AI built on a state-of-the-art language model actually delivered when it tried to read live retailer pages end-to-end. The reason is built into the technology: the AI starts each page from scratch, its knowledge frozen at the moment its training ended, and nothing it learns from your catalog carries forward to the next page. The same ceiling shows up in every published academic benchmark Part 2 cited — WebArena, OSWorld, the rest. The limit is not a bug we could have engineered our way past. It is a property of how this kind of AI works.

One detail from that experiment matters more now than it did then. The two cart anomalies we designed the agent to catch were a **stockout** and a **price mismatch where the cart price exceeds the PDP's**. We chose them before we knew what we would build next. They turn out to be exactly the two states that machine judges punish and that paid traffic dies on. The experiment failed. Its target survived.

What we built

Re-judging is what fails. The AI we ran in Part 2 re-guessed what every page meant, every time, because that is how this kind of AI works. A person re-reads until attention collapses — and at tens of thousands of pages, no team is large or attentive enough to read each page by hand. So we removed the repetition itself.

SHORA's engine records the **template** of a retailer's page once — what kind of page it is — not the individual pages themselves. A person records it once, with full attention, on a live page exactly as a shopper sees it; the engine then replays that recording across every product page on the website, no matter how many products it sells. Replay is mechanical, not interpretive —

the same template, read the same way, as often as your operations need, with no AI doing any thinking when the engine reads your pages and no human reviewer in the loop.

The template recording survives **content changes** — a price changes, a stock level changes, a new product is added, a promo banner appears after the initial page load, an A/B test is running on the page — without re-recording. It also survives **redesigns** — a new look, new colors, new layout, a new framework under the hood. None of these trigger a re-record, because the recording is bound to the structural intent of the page, not to its surface appearance. (How the binding holds across these mutations is the part we keep.)

There is exactly one event that does require a new recording: a **change to the structural intent itself**. If the retailer re-architects what its product page is — the price stops being a price and becomes "starting from X," add-to-cart becomes "request a quote" — the existing recording fails loudly, with evidence attached, on the very first replay. You pay one new recording for that template, and replays resume across the whole website. That is the honest cost of the architecture: you pay it once, when the page becomes a fundamentally different page.

Here is how the engine answers each of Part 2's failure modes:

- **Why our engine does not make the AI's mistakes.** A person makes the careful judgment about what a page is — once per template, covering every product the site sells — and the engine then reads it mechanically forever. The AI re-guessed it on every page; the engine never does.
- **What happens when something on a page actually changes.** The engine does not improvise. If the page changes in a way that matters, replay stops and tells you, with the page in front of you as evidence. The AI's failure was the opposite — it kept producing answers, confidently, and you only learned they were wrong from a downstream symptom you might not catch.
- **The one failure type a money decision cannot survive.** Silent, confident, and wrong. That is the failure pattern the AI in Part 2 fell into — wrong outputs that compounded across the journey until something downstream broke, which is how the completion rate ended at 9% instead of the 80% bar we had committed to. The engine cannot produce that failure — when it cannot read a page correctly, it stops and says so. We did not build a system that never breaks; nobody honest can claim that. We built a system that cannot lie.
- **The cost difference.** Per completed web task, running the engine is about one-tenth the cost of running an AI agent over the same catalog. That is the difference between watching your full catalog at the cadence your operations actually need, and being forced to sample a slice of it because the AI's unit cost makes full coverage prohibitive.

Why now: three forces

The judges are multiplying. Merchant Center enforcement already audits page-versus-feed-versus-checkout agreement per item, with preemptive disapproval and account-level review as the penalty schedule. Agentic buyers extend the bench: the moment an agent transacts on a buyer's behalf through Visa's Trusted Agent Protocol or its peers, "purchasable by a machine" stops being an SEO nicety and becomes a revenue precondition.

Paid media went black-box, and the page is the last surface you own. The French digital advertising market reached €12.4 billion in 2025, up 11% year-over-year, with eight platforms capturing 76% of total spend and 83% of the year's growth. At the unit level, the Contentsquare 2025 Digital Experience Benchmark — drawn from 90+ billion sessions across 6,000 websites — reports the *cost of an online visit up 9% year-over-year and 19% over two years*, against an overall conversion rate down 6.2% year-over-year (a 7.4% drop for new customers specifically); brands that leaned harder into paid social saw conversions fall 10.6% year-over-year on that traffic. The two indicators triangulate: more euros are flowing into the same auctions, each visit is more expensive to win, and fewer of those visits convert. The industry argues about causes — auction inflation, signal loss, soft demand — and the honest answer is that the aggregate cannot be decomposed from a dashboard. But one slice of it is fully auditable: how much of the traffic you are paying for lands on products that cannot be bought. With bidding delegated to Performance Max and Advantage+, the destination state is the one variable still entirely in the retailer's hands — and the least instrumented.

Europe blinded its own analytics. Consent banners structurally amputated tag-based measurement: when "Reject All" appears on equal footing with "Accept All," roughly 60% of EU users reject — so the sessions you most need to understand are heavily the ones you are no longer allowed to watch from inside your own site. Server-side crawlers that skip the banner go too far the other way: accepting the privacy policies reveals up to 70 additional trackers the un-consenting view never observes — neither view matches what a real buyer sees. The right instrument is neither: it visits your store the way a real shopper does, meets the banner, makes a consent decision, and records everything the shopper sees from there — including the differences between the accepted and rejected paths. In the EU, this is not a loophole. It is the only way to read your full storefront the way your customers actually see it after their consent decision lands.

Three forces, one conclusion: **the buyable state of promoted products has become the highest-stakes, least-instrumented surface in commerce.**

If you act on competitor data, your tool's drift is your loss

If you make decisions from competitor data — price-matching, promo-matching, stock-monitoring — read this twice.

Every competitive comparison is a subtraction: your price minus theirs, your availability minus theirs, your promo against theirs. Subtractions amplify error. A tool that misreads your competitor's price by a few percent — because their page redesigned, because their stockout indicator changed, because a bot challenge got mistaken for an empty cart — gives you a comparison that is wrong in ways you cannot measure. You don't get a noisy comparison. You get a confidently wrong one, with no signal that anything broke.

> Where a machine judges your store record by record and your media budget rides on the comparison, an instrument that reads the same way every time stops being an engineering preference. It becomes the bar your data has to clear before anyone signs off on it.

There is a second reason this matters — the one a procurement team notices. A system that produces the same output from the same input is something your procurement team can specify, audit, and put under an SLA: **signed for**. Part 2 closed on the fact that neither the LLM nor the human delivers a measurement layer a buyer can sign for. This one does.

What this lets you measure that nothing else does

Put the four problems above next to each other — machine judges that punish mismatches, ad budgets dying on broken products, competitor failures you cannot see, tools that drift silently — and the product almost designs itself: a measurement layer for retailers who buy traffic, built on a recording engine that does not drift.

Concretely, running the engine across your promoted catalog produces three numbers nobody reliably has today:

1. **Your wasted-spend number.** Every euro of paid traffic you are spending on products shoppers cannot buy — out of stock behind a stale "in stock" badge, cart price higher than the advertised price, the promoted variant rejected at checkout. Not modeled. Replayed, time-stamped, evidence attached so your teams can reproduce and audit.
2. **Your competitor's outage windows.** When a rival's promoted bestseller goes unbuyable, that rival drops out of the consideration set for those queries — the shoppers do not vanish, they redistribute across whoever is still standing, and your listing converts a higher share of them because a direct substitute just left the shelf. And if the rival's SKU is disapproved rather than merely out of stock, it leaves the ad auction too: fewer bidders on those queries, a lower cost per click for you. The catch is that the window is only worth bidding into if you can see it while it is still open and trust the data enough to move spend inside it. The subtraction in the section above is what makes that trust possible; a drifting tool would hand you false windows, and you would burn budget chasing ghosts.

3. **Feed-vs-page-vs-checkout agreement, before the clock starts.** This is the 3:47 a.m. problem, caught on your terms instead of Google's. The engine reads each page the way the crawler reads it — the raw HTML the server returns before any JavaScript runs — and the way a shopper reads it — the fully-rendered page — and it compares both against your feed and your checkout. Where the crawler-view price and the shopper-view price diverge is exactly the gap that triggers a disapproval; the engine surfaces those SKUs, with the page captured as evidence, every hour you ask it to, *before* Google's 28-day clock starts — not after, when the listing is already gone and the page that caused it is unreproducible.

"Can't we just have AI write the extraction tool for us?"

AI-written automation is still automation that breaks when pages change — you have swapped who wrote the code, not what happens when it fails. The breaks you catch are cheap; the silent ones drain your media budget while everyone trusts the dashboard. Catching them is a vigilance task — the exact kind of repetitive monitoring work Part 2 showed humans measurably fail at.

What you get if you run this on your catalog

If you run paid media or e-commerce inside a retailer that buys traffic, you have now seen the three numbers. What you get is those numbers for your catalog and your top competitors, with evidence your teams can reproduce and audit. What we ask is read access to GA4, Google Ads, and Merchant Center diagnostics — so we compute your wasted-spend number from your data, not from our assumptions.

Timing matters most in the quarter you are about to plan. Every hour a competitor's bestseller sits dead between Black Friday and the holidays is the cheapest demand of your year — but only if your measurement is already running, and already trusted, before the season starts. The person who walks into the next budget review with the wasted-spend number does not argue for budget; the number does.

Every month this stays unmeasured, paid traffic keeps landing on products nobody can buy while your competitors' outage windows close uncaptured. If your paid traffic never lands on a dead end and your competitors never stock out, you do not need us. Otherwise, it is thirty minutes with your catalog.

Built at INRIA. No model in the data path. No reviewer in the loop. No exceptions.
