

Free for humans, metered for machines.

Cloudflare built a toll booth for AI crawlers and left the pricing logic empty on purpose. Eleven businesses that could grow in that gap, what each one actually requires to work, and the specific reason each one might not.

BY DAVE BROWN

For about thirty years the commercial internet ran on one bargain. Humans read for free, publishers got traffic, advertisers paid for access to the humans. Nobody signed it, everybody honored it, and the crawler was the least controversial participant in the whole arrangement, because a crawler that indexed you also sent people back.

AI crawlers broke the return leg. They read enormously and refer almost nothing, and the product built on top of what they read frequently answers the reader's question without the reader ever arriving. That is not a moral complaint; it is an accounting one. The cost side of serving a publisher's content went up and the revenue side went away.

The interesting part isn't that a toll booth exists. It's that nobody has decided what anything costs.

Cloudflare's answer is **Pay Per Crawl**, which resurrects HTTP status code 402 — "Payment Required," a code that has sat unused in the spec since 1997. A crawler either arrives with payment intent in its headers and gets the page, or it gets a 402 with a price attached and decides for itself. Cloudflare handles the money. Publishers set allow, block, or charge.

The mechanism is genuinely new. What follows is a set of businesses that could exist around it, and a serious effort not to oversell any of them. Every idea in here gets a matching paragraph on why it might fail, because the honest state of this market is that a real primitive shipped into a demand environment nobody has measured yet.

402

the HTTP code doing the work, unused for nearly thirty years

~20%

of global web traffic sits behind the company running the toll booth

50,000:1

worst reported crawl-to-referral ratio for a major AI bot

11

businesses here, each with the reason it might not work

One framing note before the list. The temptation is to look at pay-per-crawl and see Stripe in 2011: a payment rail that will grow a hundred companies on top of it. That analogy is useful and it is also doing a lot of load-bearing work it hasn't earned. Stripe launched into a market where the transactions already existed and were merely painful. Pay-per-crawl launched into a market where nobody has yet demonstrated that the buyer wants to transact at all. Hold both of those thoughts.

—— WHAT'S IN HERE

- 01 What actually shipped. Separating the live mechanism from the pitch deck around it.
 - 02 The empty hook. Why the pricing layer is the interesting part, and where that argument strains.
 - 03 Sell side. Five businesses aimed at publishers trying to earn from machines.
 - 04 Buy side. Three aimed at the companies that suddenly have a new cost line.
 - 05 Trust and format. Two that only matter if the market gets big enough to cheat.
 - 06 The strange one. A publisher whose actual customer is the bot.
 - 07 What has to be true. Four conditions, and the case against each idea surviving.
 - 08 Signposts and a cheap test. What to watch, and thirty days that beat a business plan.
-

A real mechanism shipped. The market around it is still hypothetical.

Most writing about this shift blurs together things that are running in production, things that are in closed beta, and things that are somebody's guess. Worth pulling apart before you build on any of it.

Here is the sequence, with the confidence level attached to each item, because the confidence level is the part that changes what you should do about it.

WHEN	WHAT HAPPENED	HOW SOLID
July 2025	Cloudflare launches Pay Per Crawl in private beta. HTTP 402, publisher sets one flat per-request price for the whole domain, Cloudflare acts as merchant of record.	Announced publicly, still closed beta. Participation numbers not published.
Through 2026	Advanced configuration: exempt URL patterns from charging, and set a per-URL price by returning a crawler-price response header from your origin or computing it in a Worker.	This is the load-bearing detail for most of what follows.
2026	A Discovery API lets verified crawlers find domains participating in pay-per-crawl. An x402 payment-gated proxy template ships as a reference implementation.	Real, and worth noticing: two of the "startup ideas" people pitch are already Cloudflare templates.
September 2026	Default behavior changes. Crawlers that blend search, training, and agent use get blocked by default on pages carrying ads, for new customers, new sites, and existing free accounts.	The forcing function. Defaults move markets far harder than features do.
Ongoing	Competing rails: the RSL licensing standard, IETF work on machine-readable AI preferences, other CDNs, and direct bulk licensing contracts between labs and large publishers.	The reason "Cloudflare owns this" is not a safe assumption in either direction.

Two numbers get quoted constantly in coverage of this, and both deserve a caveat before you build a thesis on them.

- **“Bots are more than half of web traffic.”** True in the broad sense, and mostly irrelevant here. That figure counts every automated request on the internet — scrapers, monitors, vulnerability scanners, credential stuffers, ad verification, uptime checks. The slice that is an AI crawler with a budget and a willingness to pay is much smaller, and nobody outside the labs and the CDNs knows how much smaller.
- **“Crawl-to-referral ratios run to 50,000:1.”** Also broadly true, and it proves the old bargain is dead. It does not prove the new one pays. A ratio is not a revenue number. Fifty thousand crawls at a tenth of a cent is fifty dollars, and that is the entire argument of the skeptics in one sentence.

THE ONE FACT THAT MATTERS MOST

Cloudflare built the toll booth, the payment rails, and the merchant-of-record relationship — and then deliberately left price as an empty hook. `crawler-price` is a header they are happy to let somebody else compute. That is an open socket in front of a fifth of the web's traffic, and almost nothing is plugged into it.

That is the real opening. Everything in Part three flows from it. It is also, as Part seven will argue at some length, an opening that assumes the socket ever carries meaningful current.

PART TWO · THE THESIS

Every payment rail eventually grows a pricing layer.

The pattern is consistent enough to be worth betting on, and specific enough to be worth checking whether the conditions that produced it are actually present here.

Google shipped AdWords with manual bids. Within a few years, bid management, attribution, and yield optimization were their own industry, and some of those companies were worth more than the ad products they optimized. Airlines shipped seats at a fixed fare and grew revenue management departments that now decide the price of the aircraft. Stripe made the transaction easy and did not eliminate a single company doing subscriptions, dunning, fraud scoring, tax, or analytics. It created them.

The mechanism is always the same. A platform ships crude monetization because crude monetization is what you can ship. The market then discovers that the margin lives in the pricing decision, not the payment. Specialists move into the pricing decision. Eventually the specialists are load-bearing.

Pay-per-crawl is unambiguously at the crude stage. A flat per-request price across a whole domain is the same as charging one rate for a Super Bowl spot and a 3 a.m. infomercial. Today's investigative piece that took three reporters six months and a 2013 high school basketball recap are not worth the same to a model, and pricing them identically is leaving the entire interesting problem on the floor.

The business isn't "we'll let you price your pages." Cloudflare already does that. It's "we'll tell Cloudflare what every page should cost."

A pricing engine would look at signals a flat fee ignores entirely: freshness and velocity, whether you are the only source and for how long, existing ad yield on the page — charge machines /less where humans already monetize well and more where they don't — depth and originality, byline authority, how scarce the topic is in the training corpus, and crucially which crawler is asking, because a frontier training run and a comparison-shopping agent have wildly different willingness to pay for the same paragraph.

Now the part most versions of this argument skip. The AdWords analogy holds only because three conditions were true there: there were millions of buyers, the buyers were demonstrably price sensitive, and there was an auction producing a continuous demand curve. Pay-per-crawl currently has perhaps a dozen buyers that matter, no public evidence of price sensitivity, and a take-it-or-leave-it posted price rather than an auction. A dozen buyers is not a market, it is a negotiation.

That does not kill the thesis. It does mean the honest version of it is conditional, and the conditions are worth stating up front rather than in a disclaimer at the end. They are in Part seven.

Five businesses aimed at publishers.

These share one buyer — a publisher trying to convert machine traffic into money — and therefore one structural weakness. Publishers are broke, slow to buy, and have been sold four consecutive waves of technology that were going to save them.

Each entry below has the same shape: what it is, who buys it, and the specific way it dies. The last part is not decoration. For most of these the failure mode is more likely than the success case, and knowing which one you are betting against is the whole job.

IDEA 01

The pricing engine.

The anchor. A service that sits behind the crawler-price hook and answers one question for every URL, continuously: what should this machine pay to read this right now? Start with rules a publisher would agree with on sight — new is worth more than old, original reporting is worth more than a wire rewrite, a page with no ads on it should charge machines more than a page with good ad yield. Then add experimentation, abandonment measurement, elasticity estimates, and demand forecasting.

You need to own none of the hard infrastructure. Not payments, not crawler authentication, not the edge. Cloudflare built exactly the primitive this requires and left it addressable. You are airline revenue management, except instead of optimizing seat 14C from Boston to Chicago you are optimizing article #8,174 against a crawler with a budget.

WHO BUYS

Mid-size and large publishers with archives worth differentiating

REVENUE MODEL

SaaS plus a share of incremental crawl revenue

PLATFORM RISK

Medium — adjacent to the rail, but it is a data business, not a feature

THE RISK

You would be optimizing a variable nobody has shown is elastic. If crawlers treat price as a binary — pay anything under my ceiling, refuse everything above it — there is no yield curve to manage and the entire product collapses to a rules engine a competent publisher writes in a Worker in an afternoon. Before writing a line of code, the question to answer is whether crawler behavior actually changes between \$0.001 and \$0.01. If it doesn't, there is no company here.

IDEA 02

Analytics for a revenue line that didn't exist last year.

Publishers measure page views, sessions, ad impressions, subscription conversions, referral sources. None of that describes machine revenue. A bot-funded page needs a different stack entirely: revenue per URL, per crawler, per section, per author, per content age. Bot RPM. Revenue decay curves after publication. Crawl abandonment when price changes.

The version that matters isn't the dashboard, it's the sentence the dashboard eventually produces for an editor: *your technology desk earns four times more machine revenue per article than entertainment does*. That is the moment analytics stops being reporting and starts changing what gets commissioned. It is also the moment it becomes hard to displace.

WHO BUYS	REVENUE MODEL	PLATFORM RISK
Publisher revenue ops, then editorial leadership	Seat-based SaaS, upsold from the pricing engine	High — a dashboard is the cheapest retention feature a platform can ship

THE RISK

Analytics on a revenue line worth forty dollars a month is a hobby, not a purchase. This product has no independent right to exist until crawl revenue is material to somebody's P&L. Worse, Cloudflare holds the underlying data and has every incentive to ship the dashboard free, because the dashboard is what makes publishers keep the feature turned on. Building analytics on top of somebody else's data, for a metric they also want to report, is a well-documented way to lose.

IDEA 03

Crawl yield optimization, or SEO for machines.

An entire twenty-five-year-old discipline exists to make content findable and attractive to a human via a search ranking. The mirror discipline is making content legible, complete, and genuinely worth paying for to a machine: clean structure, resolvable entities, explicit dates and provenance, tables that survive extraction, internal linking that an agent doing multi-hop reasoning can actually follow, formats that don't require a headless browser and 4,000 lines of CSS to parse.

The consultancy version bills a retainer starting next week. The software version is an audit tool that scores a site's machine-readability and tells you what to fix. Historically the consultancy version funds the software version, and the people who did that in 2004 for search did fine.

WHO BUYS

Publishers, then anyone whose content is an asset

REVENUE MODEL

Retainer first, product second

PLATFORM RISK

Low — services are unpleasant enough that platforms leave them alone

THE RISK

The optimization target is set by counterparties who publish nothing. Google, for all its faults, documented what it wanted; frontier labs document almost nothing about how they value a page, and they change it without notice. You would be selling advice about a black box you cannot query. And the incentive gradient here runs somewhere ugly: the fastest way to raise crawl yield is to fragment articles into more billable requests and pad thin pages, which is fraud with better branding and gets audited out the moment anyone is checking. Sell the honest version and you are competing against people selling the other one, cheaper.

IDEA 04

The 402 layer for the four fifths of the web that isn't behind Cloudflare.

Roughly 80% of the web sits somewhere else and still has the same crawler problem, minus the tooling. There are two shapes here. The middleware version packages bot detection, 402 handling, and payment collection as a WordPress plugin, an edge function, an nginx module, a Shopify app — distribution through the platforms the long tail already runs on. The broker version doesn't ask anyone to move: you route only crawler traffic through infrastructure that negotiates payment on the publisher's behalf and remits the revenue net of a cut. "Keep your stack, get the toll booth."

The broker shape has a good precedent. Most merchants never touch the card networks directly; they use a payment facilitator that aggregates them into something the network will talk to. Individual small publishers are beneath the notice of a frontier lab's procurement process. Two hundred thousand of them behind one endpoint are not.

WHO BUYS

The long tail, via plugin marketplaces and hosts

REVENUE MODEL

Percentage of collected crawl revenue

PLATFORM RISK

Medium — Cloudflare already publishes a reference proxy for exactly this

THE RISK

The 402 is the easy part, and Cloudflare gives away a template that does it. The hard parts are bot verification you can trust at the edge and being merchant of record for millions of sub-cent cross-border transactions, which is a compliance and accounting problem long before it is a revenue stream. There is also a brutal cold-start issue: a lone site with a toll booth on it does not get paid, it gets skipped. Nothing about this works until crawlers have a reason to look for your endpoint, which means you need publisher density before you have a single dollar of GMV.

IDEA 05

A licensing collective for the long tail.

A niche site is a rounding error to a frontier lab and has precisely zero negotiating leverage. Forty thousand niche sites in the same vertical are a corpus, and a corpus can be licensed. Build the ASCAP of crawl: aggregate the rights, negotiate one deal, meter usage, distribute royalties on some defensible basis.

This is also the direct answer to the strongest criticism of the whole model — that per-crawl pennies cannot replace reader revenue for a small publisher. That criticism is correct at the level of the individual site and much weaker at the level of the pool. Aggregation is the mechanism by which pennies become a check, and it is how every previous rights market for small creators eventually got built.

WHO BUYS

Labs buy the license;
publishers join the pool

REVENUE MODEL

Administrative percentage of
distributed royalties

PLATFORM RISK

Low — Cloudflare structurally
will not represent publishers
against buyers

THE RISK

Collective price-setting by competitors is antitrust exposure, not a business model detail. ASCAP and BMI operate under federal consent decrees for exactly this reason, and the recent news-bargaining regimes in Australia and the EU exist because ordinary competition law does not permit publishers to set a price together without a statutory carve-out. Any credible version of this needs real counsel before it needs a landing page. The second problem is quieter: long-tail content is, almost by definition, the least scarce content on the internet. Aggregating forty thousand sites that all cover the same commodity topics can aggregate to zero leverage, because the buyer can walk away from all of them at once.

Three businesses aimed at the companies now facing a bill.

Every transaction has two ends, and the underserved end is usually the one that has to pay. These are structurally more interesting than the sell-side ideas, because the buyers are solvent and because Cloudflare will not build them.

Cloudflare represents publishers. That is the whole positioning of the product, and it means an entire category of tooling — anything that helps a crawler pay less, or decide not to pay — is something the platform is structurally prevented from shipping. That is the most durable moat in this entire document, and it belongs to the buy side.

IDEA 06

Crawl FinOps.

An AI company that crawls at scale just acquired a cost line that did not exist eighteen months ago, sitting next to compute and inference, with no tooling around it whatsoever. Today they optimize GPU hours obsessively and content acquisition not at all, because until recently content acquisition was free.

What they will need looks exactly like cloud FinOps did in 2016. Deduplication so you never pay twice for a URL you already hold. Per-crawler price ceilings and budget caps. Shared caches across teams so the retrieval group and the training group aren't buying the same archive separately. Spend attribution by product and by query. And the decision layer underneath all of it: is this page worth \$0.003, \$0.03, or \$3.00 for this purpose, right now? Policies like *refresh financial data every ten minutes and restaurant reviews every seven days* are trivial to state and nontrivial to enforce across a crawling fleet.

WHO BUYS

Labs, agent startups, RAG-heavy products, data teams

REVENUE MODEL

Platform fee, or a share of demonstrated savings

PLATFORM RISK

Low from Cloudflare. High from the customer building it themselves.

THE RISK

The spend has to exist before the tooling can. If labs respond to tolls by routing around them, signing bulk licenses, or simply crawling less, this cost line never gets large enough to justify a vendor. And the customers here are the most build-versus-buy-hostile buyers on earth: they have world-class infrastructure engineers, this touches their cost structure and their data pipeline, and “which sources are we paying for” is competitively sensitive in a way that makes external tooling a hard sell. The realistic path is probably to start as the neutral dedup-and-cache layer between many buyers, where being external is the feature, rather than as a dashboard any of them could write.

IDEA 07

Demand intelligence, and the clearing price nobody publishes.

Publishers spend enormous sums to learn what humans want — Trends, Semrush, Ahrefs, keyword research, social listening. There is no equivalent for what machines want, and machine demand is a fundamentally different signal. It is not “what are people searching for,” it is “what are agents repeatedly trying and failing to find.”

Two products live here and they reinforce each other. The first is a demand feed: requests for commercial HVAC pricing up sharply this month, heavy agent interest in regional zoning records with almost no authoritative source available. That is not analytics, it is a content opportunity detector, and it tells a publisher what to commission. The second is the benchmark: what are crawlers actually paying, by vertical, by content type, by buyer? Nobody knows, and the two parties who do know have no incentive to say. Whoever aggregates that becomes the reference rate for the market, which is a position with genuine network effects — the benchmark improves with every participant.

WHO BUYS

Publishers, agencies, and eventually both sides of the trade

REVENUE MODEL

Free appraisal tool as the wedge, subscription data product as the business

PLATFORM RISK

Low — cross-platform neutrality is the product

THE RISK

This is the coldest cold start in the document. The data is worthless below a certain scale and you cannot reach that scale without data. Worse, the incentives are asymmetric: published clearing prices help buyers negotiate down at least as much as they help sellers price up, so the sell side may not contribute even when asked nicely. Every benchmark business in history solved this the same way — by being credibly independent for years while barely making money — and that is a long time to be unprofitable in a market that might not materialize.

The licensed cache, and the marketplace above it.

Pay-per-crawl charges again when a crawler retrieves the same unchanged content again. Buyers do not want that and publishers do not benefit from serving it. That gap is a business: a publisher-approved cache holding one canonical licensed copy, distributed to buyers, invalidated automatically when the publisher updates the article. The publisher sheds infrastructure load, the buyer stops paying repeatedly for a page that hasn't changed, the intermediary takes a cut. Less a CDN than a content wholesaler.

One layer up sits the marketplace. Today an agent discovers a site and then discovers whether it can afford it. Reverse the order: let a buyer query available inventory before crawling anything — *authoritative sources on commercial construction wages in Ohio, published in the last thirty days* — and get back sources with freshness, reputation, price, and license terms attached. Purchase what maximizes confidence inside a budget. Cloudflare's Discovery API gestures at this at the domain level; the real version is cross-provider and works at the level of the content.

WHO BUYS	REVENUE MODEL	PLATFORM RISK
Buy side pays; sell side supplies	Transaction fee, tiered access windows, cache rights	High for the cache, medium for the marketplace

THE RISK

The cache is arbitrage against a pricing inefficiency the platform can close with one line of product: a cache-hit discount, or a flat subscription tier, and the spread is gone. It is also not really an infrastructure position — you are holding and redistributing other people's copyrighted work, which is a rights posture, and one publisher who decides your license didn't cover redistribution is an existential problem rather than a support ticket. The marketplace has the ordinary two-sided cold start on top of that, made worse by the fact that the small number of buyers can credibly threaten to go around you and deal direct, which is exactly what they have been doing with large publishers already.

Two that only matter once there is enough money to be worth cheating for.

Which is a real caveat and also a real signal: if these ever become urgent, the market got big. Building them early is a bet on timing more than on the idea.

IDEA 09

Identity, reputation, and the audit that runs both ways.

Every ad market grows a fraud market within about eighteen months and an audit layer shortly after. Here the fraud runs in both directions, which is unusual and makes neutrality the actual product.

Against buyers: are publishers inflating crawl counts, splitting articles to multiply billable requests, serving thin pages at premium prices? Against sellers: are crawlers spoofing user agents, laundering requests through residential proxies, running under a hundred identities to avoid the toll? And underneath both, the reputation question — does this client pay reliably, respect caching rules, honor the license terms it agreed to? A publisher would happily configure instant access for a trusted crawler, prepayment for an unknown one, and a block for a known abuser, if something existed to tell them which was which.

There is a specific integrity problem worth naming separately, because it is the one that could poison the whole market. Once machines pay and humans don't, publishers acquire a brand-new incentive to serve machines something different from what humans see: stripped text to save bandwidth, keyword filler, or deliberately corrupted content. Verifying that a paying crawler receives what a human reader receives is the audit bureau of circulation for the machine web, and somebody has to be it. The dark twin of that business — poisoning as a service — will also exist. Worth naming as a market force; not worth building.

WHO BUYS

Both sides, which is the point

REVENUE MODEL

Certification and per-verification fees

PLATFORM RISK

High on identity, low on independent audit

THE RISK

The identity half is being standardized for free. Verified bot programs, Web Bot Auth, and IETF work on signed HTTP requests are all moving toward cryptographic crawler identity as infrastructure rather than as a product, and competing with a standard is a bad trade. The audit half has the older problem every rating agency has: somebody pays you, and it is usually one side. Neutrality is easy to claim and structurally difficult to hold, and the moment your revenue concentrates on the buy side your certifications stop meaning anything to the sell side. Also, plainly: fraud detection is only a business after there is real money to defraud. Today there isn't.

IDEA 10

Sell the API, not the HTML.

The highest-leverage move for a publisher isn't charging more per scraped page. It is making scraping the worse option. Bots do not want your navigation, your newsletter modal, your cookie banner, or your carousel; they want the text, the entities, the tables, the dates, the byline, and the provenance. Give them that as a product.

The human gets `article.html`. The machine gets a clean structured representation — an MCP server, a licensed feed, precomputed embeddings, real terms attached, no parsing required. Then price it above the crawl rate and let the buyer choose convenience. *Don't crawl me, query me*. You can even sell it in tiers: raw text cheap, structured with extracted entities more, structured plus revision history and source metadata more again. The business is doing this conversion for the overwhelming majority of publishers who will never build it themselves.

Taken further, this is a CMS thesis. Twenty years of publishing software was built around the web page. A CMS built around information products sold simultaneously to two very different customers would put machine metadata, provenance, crawler permissions, pricing, license terms, and freshness windows in the same editor as the headline. That is a much bigger and much slower company than the conversion-service version.

WHO BUYS

Publishers with archives;
anyone sitting on
structured data

REVENUE MODEL

Setup fee plus revenue share
on the feed

PLATFORM RISK

Low — this has nothing to do
with the toll booth

THE RISK

This is the idea most likely to be real, and it is a services business wearing a software costume. Every implementation is bespoke, margins behave accordingly, and the buyer is a publisher with a hiring freeze. The deeper issue is that when a large buyer genuinely wants your archive, they would rather sign a licensing contract with a lawyer than integrate with your feed — the bulk deals already happening between labs and major publishers are evidence that the preferred transaction shape at the top of the market is a contract, not an endpoint. That leaves you selling infrastructure to the tail, which is the segment with the least money.

Four more worth naming, thinly.

Each of these is a real gap and none of them is obviously a company yet. Listed because if the market grows, these are where the second wave forms.

Proof of humanity as a service. If ad-bearing pages become the trigger for default blocking, the cost of a false positive is a real reader hitting a toll booth. Somebody has to make sure humans never do.

Agent checkout for consumers. When your personal agent reads on your behalf, whose card is on file? There is a wallet-and-consent problem here that nobody has built, and it is the one piece that could make this market consumer-scale rather than enterprise-scale.

Revenue recognition and tax tooling. Millions of international sub-cent transactions are an accounting problem well before they are a revenue stream. Unglamorous, mandatory, and historically a good business.

Rights resolution. Paying to retrieve a page does not tell you what you may do with it — quote, summarize, cache for a day, embed, retain, fine-tune, train. Those permissions need to become machine-readable, and a crawl worth a cent and a training license worth fifty dollars are not the same transaction at all. This is a standards fight before it is a product.

A publication whose actual customer is the bot.

Every other idea here adds machine monetization to something that already exists. This one starts from the machine and works backward, which makes it either the most interesting item on the list or the most reckless.

Don't retrofit a publisher. Start one where machines are the paying audience from day one. Humans read everything free — no paywall, possibly no ads. The entire editorial operation is funded by machine consumption.

The workflow is genuinely different from journalism as practiced. You study machine demand to find what agents repeatedly need and cannot find: municipal zoning records, regional construction costs, obscure regulatory updates, industrial equipment pricing, structured local government proceedings. Then you commission authoritative original work in that gap, structure it perfectly for machine consumption, and charge for access. You are not competing for attention. You are competing to be the best available answer in a category where the current best available answer is nothing.

You're not building an audience. You're building the source a model reaches for when it has no good option.

What makes this compelling is that it inverts the constraint that has strangled digital publishing for two decades. You no longer need scale, virality, or an SEO strategy. You need to be authoritative and unique in a narrow domain that machines demonstrably need and humans were never going to make profitable on ad impressions. There are thousands of those domains.

THE RISK, AND IT IS SEVERE

You would be building a business whose only revenue channel is a closed beta run by a single vendor, selling to perhaps four to eight buyers with genuine monopsony power. Single-buyer markets do not have prices, they have terms, and the terms are set by the buyer. If the labs decide your vertical isn't worth paying for, you have no consumer brand, no subscriber list, and no advertiser relationships to fall back on — you have a website nobody visits and a cost base.

The survivable version hedges: pick a vertical where the same research is sellable to human professionals too, so the machine revenue is upside on a business that already works rather than the whole thesis. That is a less exciting company and a considerably more likely one.

PART SEVEN · WHAT HAS TO BE TRUE

Four conditions, and none of them is settled.

Ignoring these makes the whole thesis read like a pitch deck. Handling them is the difference between a market map and a guess with numbers on it.

Everything above rests on four assumptions. They are load-bearing in the sense that if any one of them fails, most of the list fails with it.

One. Enough of the web has to charge that routing around is expensive. A 402 the crawler declines is a block with extra steps. Content is duplicated, syndicated, aggregated, and reposted constantly; if the same information is free three hops away, the toll booth collects nothing and merely removes you from the answer. This condition favors genuinely scarce content and punishes everything else, which is an uncomfortable finding for the long-tail ideas.

Two. Crawlers have to be price sensitive, not binary. Yield management requires a demand curve. If buyers set one ceiling and treat everything under it as free and everything over it as blocked, there is no curve, and the pricing engine — the anchor idea of this entire document — degrades into an if-statement. This is the single most testable assumption on the list and the one with the least public evidence behind it.

Three. Payment has to buy rights, not just bytes. A cent for retrieval does not settle whether the content can be used for training, and the litigation determining what a lab may do with material it lawfully accessed is unresolved. Until that clarifies, the largest buyers will prefer negotiated contracts with indemnities over a header that says a price. That preference pulls the top of the market away from per-request pricing entirely.

Four. The revenue has to clear a floor where tooling is worth buying. Nobody purchases a pricing engine, an analytics suite, and a compliance layer to optimize \$200 a month. Most of the sell-side businesses here need crawl revenue to be a meaningful line on a publisher's P&L before they are purchases rather than experiments, and there is currently no public evidence of a single publisher for whom that is true.

And here are the four objections that come up every time this thesis is presented, with the most honest available response to each. Two of them have decent answers. Two do not.

THE OBJECTION	THE HONEST RESPONSE
The pennies problem. Per-crawl fees cannot approach ad or reader revenue, and the platform sets the payout terms with no room to negotiate.	Largely correct, and the rebuttal is not volume. It is aggregation (idea 05) and premium pricing on genuinely scarce content (idea 01). If your content is neither scarce nor poolable, the skeptics are right about you specifically.
Cloudflare just builds all of it. They went from flat pricing to dynamic pricing in months and they ship reference implementations of the middleware.	Fair, and it disqualifies several of these. Anything trivially adjacent to the toll booth is a feature. What survives is what they structurally will not build: buy-side tooling, neutral auditing, collective bargaining against buyers, and cross-platform benchmarks.
The crawlers simply refuse. If paying is optional and skipping is free, most will skip.	No good answer yet. This is a coordination problem the market has not solved, and the September default change is essentially an attempt to solve it by force. Whether that works is the most important open question in this whole space.

THE OBJECTION

THE HONEST RESPONSE

The old bargain is already dead anyway. At these crawl-to-referral ratios there is no traffic to protect.

True, and it is the strongest argument for publishers participating even at bad prices — but notice it is an argument about the alternative being worse, not about this being good. Those produce very different businesses.

Scored against all of that, the list separates fairly cleanly. The column that matters most is the last one.

IDEA

PLATFORM RISK

THE THING THAT HAS TO BREAK YOUR WAY

01 · Pricing engine

Medium

Crawlers turn out to be price sensitive

02 · Bot analytics

High

Crawl revenue becomes material to a P&L

03 · Crawl yield optimization

Low

Buyers publish enough for “optimization” to mean something

04 · 402 for everyone else

Medium

You reach publisher density before you run out of money

05 · Licensing collective

Low

Antitrust structure holds, and the pool has real leverage

06 · Crawl FinOps

Low / customer-build risk

The buy-side cost line gets big enough to manage

07 · Demand intelligence

Low

You survive the cold start long enough to be the benchmark

08 · Licensed cache and marketplace

High

The platform doesn’t close the spread; rights hold up

09 · Identity and audit

High on identity

There is enough money in the market to be worth defrauding

10 · Sell the API

Low

Nothing, really — this works even if pay-per-crawl doesn’t

11 · Bot-first publisher

Low

A handful of buyers stay willing to pay, indefinitely

Idea 10 is the outlier and it is worth saying plainly. Publishing your archive as a clean, machine-native, properly licensed product is good business whether or not anyone ever pays a 402. It makes you easier to cite, easier to license directly, and cheaper to serve. If you want one thing off this list that does not require the thesis to be correct, it is that one.

THE PART THAT ISN'T SPECULATIVE

Four things here are true regardless of whether pay-per-crawl works. Machine traffic is real and expensive to serve. Blocking is already deployed at scale and increasingly on by default. Direct licensing deals between labs and large publishers are real money changing hands right now. And the ad-funded model for written content is eroding for reasons that have nothing to do with any of this. A business that is useful under those four facts alone, with pay-per-crawl as upside rather than as the thesis, is a substantially better-shaped bet than one that needs the toll booth to succeed.

PART EIGHT · SIGNPOSTS AND A CHEAP TEST

You can find out whether this is real in about thirty days.

For considerably less than the cost of being wrong about it for a year. The point of this section is to replace the market map with evidence you gathered yourself.

First, the indicators. These are the things that would move this from an interesting mechanism to an actual market, roughly in the order they would have to happen.

IF YOU SEE THIS	IT MEANS
A second major CDN or host ships 402 handling	It is becoming a protocol rather than a vendor feature. This is the single most important signal on the list.
A named publisher discloses actual crawl revenue	There is a floor, and you can finally do arithmetic instead of estimates. Watch for a real number, not a press release.
A lab publishes or leaks a crawl budget or a max-price policy	Buyers are treating this as procurement. That is the precondition for every buy-side idea here.
Prices start varying by content type across publishers	Price discovery has begun and idea 01 has a customer.

IF YOU SEE THIS

IT MEANS

Crawl volumes fall measurably after the default change

Blocking works as leverage. Crawlers that route around instead means it doesn't.

Someone files the first cloaking or crawl-inflation complaint

Perversely, good news. Fraud follows money, and there is no fraud in markets worth nothing.

A regulator opens a look at default blocking or collective pricing

The market is large enough to matter, and idea 05 just got harder.

Second, the test. If you are seriously considering building one of these, this sequence produces better information than any amount of market sizing, and it fits in a month of evenings.

WEEK ONE

Read your own logs before you read anyone's report.

Pull crawler traffic on any site you control or can get access to. Separate AI crawlers from search crawlers from everything else. Count requests, bytes, repeat fetches of unchanged URLs, and which sections get hit hardest. You are answering one question: on this specific site, is machine traffic a cost problem, a revenue opportunity, or a rounding error? Most people arguing about this publicly have never done it once.

WEEK TWO

Test the assumption everything depends on.

Put a price on a slice of content and change it. Same content, different prices across comparable URLs, and watch whether crawler behavior differs. You are looking for exactly one thing: evidence of elasticity. If crawlers pay \$0.001 and \$0.01 at identical rates, or refuse both at identical rates, you have learned that idea 01 does not exist yet, which is worth far more than a quarter spent building it.

BE HONEST ABOUT WHAT THIS PROVES

A test on one site with one content type over two weeks is a signal, not a finding. Sample size is genuinely a problem here, and crawler behavior varies enormously by operator. Treat a negative result as strong evidence and a positive result as a reason to run it again somewhere else.

WEEK THREE

Talk to ten publishers and two buyers.

The publishers will tell you whether crawl revenue is a real line item or a curiosity, and how they decided their price — the answer is almost always “we guessed,” which is either your opening or your warning. The two buyers are harder to reach and worth ten times more. Ask what they actually do when they hit a 402, who owns that decision internally, and whether anyone has a budget for it. If nobody owns it, the buy-side ideas are early rather than wrong.

WEEK FOUR

Write down what would make you stop.

Before you commit, name the specific result that would kill it: no elasticity, no budget holder on the buy side, revenue below a threshold you set in advance, or the platform shipping the feature you were going to sell. Write the numbers down. The failure mode in emerging markets isn't picking the wrong idea, it's never defining what disconfirmation looks like and spending three years interpreting ambiguity favorably.

THE REFRAME WORTH KEEPING

Being early to a market is only valuable if the market arrives. Most of the money in previous platform shifts went to people who were early by about eighteen months, not five years. If your thirty days say the demand is not there yet, the correct move is usually to build the version that is useful today — idea 10 — and stay close enough to move when the signals in the table above start landing.

So: a genuinely new mechanism shipped, a default is about to change in a way that forces the question, and the pricing layer on top of it is empty. That much is real. What is not yet real is a demonstrated buyer, a clearing price, an elasticity curve, or a single public example of a publisher for whom this revenue matters.

The pattern says a pricing layer forms above every payment rail and often ends up worth more than the rail. The pattern also says most people who arrive at the very beginning of one spend their capital waiting. The useful position is the one that pays for itself while you find out which of those you are in — which, unhelpfully and accurately, is the actual answer.

ONE MORE THING

If you're sitting on crawler logs and want a second read.

Which is, admittedly, week one of the test performed by somebody who isn't you.

The most interesting thing about this whole shift is that almost nobody arguing about it has looked at real numbers, because the real numbers are sitting in server logs that nobody has bothered to segment. Ten minutes with a log file settles more of this than a week of thinkpieces, and the answer is different for every site.

So I'm doing a small number of informal reads. Send me what your crawler traffic actually looks like — volume, which bots, which sections, whether anything is being paid — and I'll tell you which of the eleven ideas above your data supports and which one it quietly rules out. No pressure; everything here is yours to use either way.

GET IN TOUCH

Email hi@davecto.com with the subject line **"Bot Economy"** and a couple of sentences about what you're seeing in your logs.

More guides like this one, for people trying to use AI without embarrassing themselves.

Weekly, plain-language breakdowns on Instagram.

[@davecto](#)

A note on sourcing: the mechanism described here — HTTP 402, the **crawler-price** header, Worker-computed pricing, the Discovery API, and the change to default blocking — comes from Cloudflare's own announcements and changelog, and product details in a closed beta move faster than any guide can. Verify current behavior against Cloudflare's documentation before building on it. The traffic figures are widely reported industry estimates rather than audited numbers, and I have tried to flag where a statistic is being asked to carry more weight than it can. The eleven ideas, the scoring, and the thirty-day test are argument rather than research: no company described here exists as far as I know, no revenue figures have been disclosed by anyone in a position to know them, and the honest summary of the whole space is that a real primitive shipped into a demand environment nobody has measured. I'd rather you ran the test than took my word for it.