

Pay-Per-Crawl Protocol Viability: Due Diligence & Stress Test Report

1. Problem Reality Check

Scale of the Pain: Publishers face a rapidly growing threat from uncompensated AI scraping. Media publishers are estimated to lose over **\$2.3 billion per year** in ad revenue due to AI bots harvesting content without return traffic ¹. New data shows **referral traffic has plummeted** as users get answers from AI models instead of clicking through to source sites ². For example, in **June 2025** Google's crawler gave 1 referral per 14 pages scraped, but OpenAI's gave **1 referral per 1,700 scrapes** and Anthropic's a shockingly **1 per 73,000** ³. In one case a major sports site got *13 million* bot visits in a month but only **600 human visits** in return ⁴ – a virtually **parasitic** exchange. This translates to direct **financial loss** (lost ad impressions, subscription conversions) and **content devaluation** as high-quality content is used to feed AI products that disintermediate publishers.

Legal and Compliance Risk: Tensions are boiling over into legal battles. **70+ lawsuits** have been filed against OpenAI, Meta, Microsoft, and others by news publishers, authors, and artists over unauthorized scraping ⁵. Major outlets like *The New York Times* and local newspapers have opted to sue AI firms for copyright damages rather than give free access ⁶. Publishers argue AI companies are violating terms of service and copyright by using content without consent ⁷. This legal uncertainty poses a **significant risk** to AI developers – potential injunctions or damages – and highlights an urgent need for a clearer framework for content usage. It's creating an **"AI consent crisis"** where neither side has a stable legal footing.

Content Value Erosion: Beyond revenue, publishers fear a long-term erosion of content value and quality. With AI-generated summaries ("**AI Overviews**") keeping readers off publisher sites, some report **40% traffic declines** from search ⁸. AI chatbots now deliver answers sourced from publishers' content but **offer little attribution or traffic back** ⁹ – effectively *devaluing original journalism*. Industry observers warn of **"digital dumping"**, where the market is flooded with AI-generated content derived from real journalism without reward to its creators ¹⁰. This undermines the incentive to produce quality content and is viewed as an **existential threat** by many publishers ¹¹ ¹².

Existing Workarounds: At least three proxy solutions are already in play as stop-gaps:

- **Blocking AI Crawlers:** The majority of publishers have resorted to blocking known AI bots via `robots.txt` or firewall rules. A Reuters Institute survey found **73% of publishers have blocked AI crawlers** as of 2024 ¹³. However, this is a blunt instrument – it sacrifices potential AI-driven reach and is increasingly ineffective (AI agents are bypassing `robots.txt` 13% of the time, up from 3% a quarter earlier ¹⁴). Cloudflare even **turned on default AI bot blocking** for new domains ¹⁵ ¹⁶, reflecting how critical this defensive move has become.

- **Direct Licensing Deals:** Some large publishers strike **one-off licensing agreements** with AI firms for training data or content use. For instance, OpenAI signed multi-year deals with *The Atlantic*, *Vox Media*, and *News Corp (Wall Street Journal)* to license archives and provide citations ¹⁷ ¹⁸ . These deals give publishers a lump-sum payment or revenue share and sometimes access to AI tech in return. However, they are limited to **big players** – only major publishers have the leverage to negotiate, and terms are opaque and non-standard ¹⁹ . Smaller outlets are left out, and even big publishers are unsure if these deals truly compensate for lost traffic.
- **Bot Paywalls & Monitoring Tools:** A new crop of solutions lets publishers meter and charge bot access. **TollBit** operates a “bot paywall” marketplace where AI scrapers are redirected and asked to pay a toll (token) for content ²⁰ . TollBit reports signing up over **2,000 publishers** (e.g. Penske Media, *Time*, Hearst) to monetize AI bot hits ²⁰ ²¹ . Similarly, **Cloudflare’s Pay-Per-Crawl** (launched mid-2025) offers a platform to charge per request or block by default through its CDN ²² ³ . These tools are early-stage but validate that both sides are exploring transactional access models instead of “allow all or block all.”

Critical vs. “Nice-to-Have”: For many content owners, this problem is **critical and urgent**, not a mere nice-to-have. The fact that nearly three-quarters of publishers are actively blocking AI (despite the risk of losing some search visibility) shows acute pain ¹³ . News/Media Alliance (an industry group) has called unchecked AI scraping an “existential threat” to journalism’s economics ¹¹ ¹² . Publishers are incurring real costs from bot traffic (bandwidth, infrastructure strain ²³) with **zero compensation**, effectively subsidizing the AI companies. While smaller sites or those less reliant on search traffic may see this as a lower priority, the **market-wide sentiment** – from major newspapers to niche blogs – is that the status quo is unsustainable. In short, the pain point is **very real** and growing, though the **willingness to act** (or pay for solutions) will vary by publisher size and reliance on organic traffic.

2. Market Sizing & Growth

TAM (Total Addressable Market): The addressable market for **permissioned AI data access** is potentially vast, encompassing all web content that AI models might seek to ingest under a paid/permissioned framework. As a proxy, consider the current uncompensated value extraction: media publishers alone lose ~\$2.3B annually to unlicensed scraping ¹ . This figure likely understates TAM, as it excludes other content verticals (forums, encyclopedias, user-generated content platforms, code repositories, etc.) that AI consumes. Another lens: OpenAI and similar spent an estimated **hundreds of millions** on building and training models – much of that value came from freely scraped data. If even a fraction of that spend converts to licensing fees under regulatory pressure, annual TAM could reach the **low billions in the near term**. By 2027, one analysis suggests **\$2B+** could flow to publishers if a pay-per-crawl model is broadly adopted by top sites ²⁴ . Longer term, as generative AI integrates into every industry, the appetite for high-quality, rights-cleared data could make this a **\$10B+ market by the end of the decade**. Key TAM assumption: that regulation or publisher coordination forces AI companies to shift from free scraping to paid access for a significant share of their training and real-time data needs.

SAM (Serviceable Available Market): In the short run, the realistically serviceable market is narrower. Early adoption will skew toward **news media, large digital publishers, and data providers** who feel the pain most. For example, the top ~10,000 domains (major news sites, reference wikis, etc.) account for a disproportionate amount of valuable training content – these could form the initial market nucleus. If each of these could generate, say, \$50K–\$200K/month via pay-per-crawl fees ²⁵ , the **serviceable market** might

be on the order of \$3–5B/year. However, achieving this depends on high adoption within that cohort. Another segment is **AI startups needing niche data** (e.g. finance, scientific content): they may be more willing to pay early ²⁶ ²⁷, representing a serviceable sub-market where demand for compliant data is urgent. We also include regions with favorable regulation (e.g. EU, where compliance demands will be highest). In sum, SAM in the next 1–3 years could be a few hundred million dollars (if focusing on early adopters and pilot programs), growing to a couple of billion as major publishers sign on.

SOM (Serviceable Obtainable Market): For an early-stage protocol startup, the obtainable slice will start small – perhaps on the order of **tens of millions in transaction volume** in the first year or two, if focused on an “atomic network” of early adopters. A realistic SOM strategy is to target a high-value niche to prove the model: for instance, **business news and financial data publishers** (who have very high-value content for AI) or a geographic/regulatory niche (e.g. EU publishers preparing for the AI Act). By constraining focus, the startup could capture, say, 50 key publishers and a handful of AI buyers in a closed loop. If each publisher in that niche could earn ~\$1M/year in fees (e.g. from specialized AI services), and the platform takes a 5–10% cut, that’s **\$2.5–5M annual revenue** to the startup in a pilot phase. Scaling beyond will require expanding to adjacent verticals and onboarding more AI companies once the concept is validated.

Growth Drivers & Tailwinds: The market is poised to grow **rapidly** due to external pressures: - **Regulatory Tailwinds:** The **EU AI Act (expected enforcement ~2026)** is a game-changer. It imposes strict requirements for data provenance, quality, and legal compliance for “high-risk AI” ²⁸. AI firms will effectively need **audit trails and consent** for data – exactly what a permissioned crawl protocol provides. Similarly, privacy laws and copyright reforms in various jurisdictions are trending toward **opt-in regimes** for data mining ²⁹, flipping the default in favor of content owners. These create a strong incentive (even necessity) for AI companies to seek licensed data sources (“compliance-in-a-box” solutions ²⁸). - **Lawsuits & Settlements:** High-profile lawsuits could accelerate adoption. A big judgment or settlement (e.g. if publishers succeed in court) would push AI players to avoid litigation by entering marketplaces. Conversely, if AI companies preempt lawsuits by striking deals (as OpenAI did with major publishers ¹⁷), it validates the paid access model. We are essentially seeing the **beginnings of an AI content licensing economy**, akin to how music streaming compelled licensing after Napster – a legal inflection point can rapidly expand the market. - **Macro AI Growth:** The sheer expansion of AI applications (search bots, assistants, domain-specific LLMs) means data needs will **explode**. Every new AI feature (from Bing’s chat to startup chatbots) that answers real-time queries needs a steady feed of fresh content (RAG). As AI uses broaden, the volume of crawls – and thus potential paid transactions – is rising exponentially (AI bot traffic rose ~87% in one quarter on TollBit’s network ³⁰). Even at pennies per request, millions of requests scaled across use cases translate to a significant market.

Potential Cliffs: Despite growth, a few factors could cap or shrink the market: - If courts **declare broad web scraping “fair use”** under copyright, it could undermine the leverage to require payment (especially in the U.S.). This seems unlikely to fully invalidate contracts, but a favorable precedent for AI firms would weaken publishers’ position. - **AI workarounds or data substitutes:** If faced with fees, AI companies might double down on synthetic data or public domain datasets to reduce reliance on paid content. For instance, using Wikipedia (donation-supported) or community-run datasets could limit demand for paid access to news sites. Similarly, if AI can use older web snapshots (Common Crawl) without fresh updates, some may opt to tolerate staleness over paying. - **Publisher fragmentation:** If no standard emerges and multiple incompatible paywall systems fragment the web, AI companies might avoid the hassle by sticking to content from the “open” part of the web. A fractured market (each publisher with its own scheme) would be less attractive to service as a whole, possibly stalling growth until standardization occurs.

On balance, the **macro trend** is pushing toward a **paid data economy** for AI. The question is not if the market will form, but how fast and who orchestrates it. Our assumptions are transparent: growth depends on regulatory pressure and collective action by publishers; absent those, the market could remain a “nice idea” with limited uptake.

3. Competitive Landscape

Current competition in the “paid crawl” or AI content licensing space is a mix of large infrastructure incumbents and specialized startups. Below are the top existing players and potential entrants, along with their status:

- **Cloudflare (Public, ~\$1B+ ARR):** *Status:* Launched “**Pay Par Crawl**” marketplace in private beta (July 2025) ²². As a CDN handling ~20% of internet traffic ¹⁵, Cloudflare can deploy this feature to thousands of sites easily. It acts as **Merchant of Record** and uses HTTP 402 responses to charge AI bots through Cloudflare accounts ³¹ ³². *Strengths:* Massive install base (millions of websites) and existing bot management tools; effectively can flip a switch to block or charge bots for any site on its network ¹⁶. Strong trust and ease-of-use for publishers (no crypto required initially ³³). *Weaknesses:* Closed ecosystem (both publishers and AI crawlers must use Cloudflare accounts) ³², which some AI companies may resist. Currently fiat-based (though exploring stablecoins) ³³, which could mean higher transaction costs and less flexibility. Also, as an incumbent, Cloudflare may be slower to embrace open standards that don’t route through its network (potential innovator’s dilemma).
- **TollBit (Startup, ~\$31M funding – Seed+\$24M Series A in 2024 ³⁴ ³⁵):** *Status:* Two-sided marketplace for publishers and AI firms, launched out of beta in late 2024 ³⁶. Has **2,000+ publisher sites** integrated (including *Time*, *Penske Media*, *Trusted Media Brands*) ³⁷ and claims early deals with AI players like OpenAI and Perplexity ³⁸ ³⁹. Provides dashboards, token-based paywall redirects, and even caching to reduce bot costs ⁴⁰ ⁴¹. *Strengths:* First-mover advantage in dedicated AI-content marketplace; focused product with real-time analytics and monetization in one package ⁴² ⁴³. Strong industry validation: funded by top VCs and even Google’s head of AI (Jeff Dean) as an angel ⁴⁴. Partnerships with CDN providers (Fastly) expand its reach via easy integration ⁴⁵ ⁴². *Weaknesses:* Being a centralized startup, it may struggle to get AI giants on board absent external pressure (so far, many AI firms still scrape for free). Publishers’ actual revenue from TollBit is unproven (market chatter wonders if it’s material) ²⁰ ⁴⁶. The model requires critical mass: if major AI bots don’t cooperate, TollBit can only offer blocking – which publishers could do themselves. Also, its approach (fiat payments, proprietary system) could be replicated by others (low switching cost if a better deal comes).
- **News/Media Alliance & ProRata (Consortium, Pre-revenue):** *Status:* In March 2025, the News/Media Alliance (a large publisher trade group) partnered with a startup **ProRata AI** to launch a content licensing framework ⁴⁷. Their product “GistAI” allows publishers to opt-in and share **50% of AI revenue** from an AI product that uses their content ⁴⁸. This is more of a **collective bargaining model** than a technical solution – a pool where AI companies (like ProRata’s own AI) share revenue with content contributors. *Strengths:* Backed by a major industry alliance, giving it clout and a large roster of news publishers. It offers publishers a turnkey way to ensure compensation (essentially a revenue share) if the AI product succeeds, without having to manage microtransactions. *Weaknesses:* It relies on voluntary participation of AI companies – essentially requiring them to join the scheme. It

currently ties publishers to ProRata's specific AI product; it's not a universal solution. If ProRata's GistAI has limited market share, this yields minimal revenue. It's also not a real-time per-call payment, but a revenue split, which may not scale to broad use cases beyond that one AI platform.

- **IAB Tech Lab – LLM Content Ingest API (Standards Body, in development):** *Status:* The IAB (Interactive Advertising Bureau) is developing an **API standard** called “LLM Content Ingest” to help publishers control and monetize how their content is accessed by AI ⁴⁹. It's in early stages and would need adoption by major AI companies to be effective. *Strengths:* As an open standard from an established body, it could gain wide support and interoperability. It might allow publishers to set terms (like pay-per-use or limits) in a uniform way, reducing integration friction for AI consumers. *Weaknesses:* Standards are slow – this could take a long time to finalize and even longer to implement. It has **no enforcement mechanism** on its own; AI companies could ignore it unless there's industry consensus or regulation backing it. Essentially, it's potential future infrastructure, not an active competitor today, but worth noting as a competing **approach** (could either complement or obviate a proprietary protocol if widely adopted).
- **Major Web Incumbents (Potential Entrants):** Not yet offering competing products, but highly relevant:
 - **Google:** As both a crawler owner and a major web platform, Google could introduce its own framework (e.g. extend Search Console or `robots.txt` with paid access controls). It has a history of **resisting pay-for-index** (e.g. fighting news link taxes), but if forced, Google might prefer building an in-house “licensing exchange” rather than cede to a third party. Its knowledge of crawling and publisher relations is unparalleled, so a late entry could quickly dominate – for instance, integrating payments into Google's crawler infrastructure or offering publishers compensation in exchange for allowing Google's AI crawlers.
 - **Amazon/AWS:** AWS could fold a similar service into its cloud offerings (it already has a marketplace for data sets). Amazon has extensive relationships with publishers via Twitch (for content) and could leverage its cloud client base to push a standard. Also, Amazon's Alexa and generative AI initiatives could drive it to support a paid content ecosystem (it recently made licensing deals via Alexa's “Rufus” AI for news content ⁵⁰).
 - **Microsoft:** Through Bing and its OpenAI partnership, Microsoft is heavily involved in AI search and could implement a pay-per-access system in Bing's crawler or Azure services. If publishers blockade Bing's chatbot, Microsoft might consider paying or building an intermediary solution to ensure its AI has premium content access – possibly by acquiring or partnering with a player like TollBit.
 - **Akamai/Other CDNs:** Following Cloudflare's lead, other CDN and web security providers (Akamai, Fastly) could introduce similar paywall features for bots. Fastly already partnered with TollBit ⁴⁵, but it could also offer its own solution if demand is clear. These incumbents have existing publisher integrations and could fast-track competing offerings.
 - **OpenAI or Meta (as platform initiatives):** While currently on the “demand” side, it's plausible a big AI firm could try to create a **self-regulatory platform** – e.g. OpenAI launching a portal where any publisher can register and get compensated for content usage (to stave off lawsuits). This would be a quasi-competitor, effectively internalizing the marketplace instead of going through a third party.

To illustrate the competitive positioning, below is a **comparative SWOT analysis** contrasting the protocol startup (“Startup”) with two key competitors (Cloudflare and TollBit) and a major potential entrant (Google):

Factor	Startup (Open Pay-Per-Crawl Protocol)	Cloudflare (Pay Per Crawl)	TollBit (AI toll marketplace)	Google (Potential Entrant)
Strengths	<p>- Truly open and decentralized standard (not locked to one CDN). Crypto micropayments enable global, instant settlements at negligible cost ⁵¹ ⁵² .</p> <p>Composable with Web3 stack (DIDs for identity, on-chain logs for audit) attracting community contributions. Not tied to a single vendor – could become an industry utility if it gains trust.</p>	<p>- Massive distribution (20% of web) and easy deployment for clients ¹⁵ .</p> <p>Full-stack solution (bot detection, enforcement, payment handling) already in place on a battle-tested network. Strong publisher trust in security and performance; one-click to enable for existing customers.</p>	<p>- Focused solution purpose-built for this problem, with head-start in data and clients. Two-sided network already in motion (2k publishers, some AI firms) giving valuable usage data and feedback loops ²¹ ³⁹ .</p> <p>Flexible approach (offers caching, analytics, and paywall) plus VC funding for growth.</p>	<p>- Unmatched scale and resources; can integrate pay-for-content into dominant platforms (Search, YouTube, etc.). Control of Chrome/Android could enforce standards at browser or OS level. Deep pockets to cut deals or subsidize costs (e.g. could pay publishers directly to undercut competitors). Owns both sides of the market to some extent (as an AI provider and web gateway).</p>

Factor	Startup (Open Pay-Per-Crawl Protocol)	Cloudflare (Pay Per Crawl)	TollBit (AI toll marketplace)	Google (Potential Entrant)
Weaknesses	<p>- Cold start problem: needs enough publishers <i>and</i> AI clients on board to prove value.
- Novel Web3 elements might deter less tech-savvy customers (crypto wallets, etc.), at least initially.
- Legal untested: relies on smart contract licenses that haven't seen court scrutiny (could spook risk-averse enterprises).
- Very small team relative to incumbents – limited sales reach and support.</p>	<p>- Centralized model: AI firms may not want to depend on a competitor's infrastructure (Cloudflare sits in the middle of their content flow).
- Potential conflict of interest: Cloudflare's core business is selling security/CDN services; fully prioritizing content payouts might cannibalize other products or upset big tech clients.
- No fine-grained pricing initially (flat fee only) ⁵³ ; less flexibility than a smart-contract approach for complex licensing terms.</p>	<p>- Platform risk: still a startup that could be bypassed by DIY deals or knocked out by an incumbent solution.
- Requires building trust on both sides – as a new intermediary, some publishers or AI companies may be wary of dependence on a small venture-backed firm.
- Scalability questions: can the centralized platform handle billions of requests? And can it enforce compliance if an AI firm refuses to participate?</p>	<p>- Not active yet: would likely enter only if forced, so might be slow until pain is acute.
- Possible antitrust scrutiny if Google tried to dominate this space (given it already has search monopoly issues ⁵⁴).
- Publisher distrust: many publishers blame Google for current woes; they might not embrace a Google-led pay solution unless no choice.
- As an AI provider, Google might prioritize its own cost and data needs over publisher profitability (conflicting incentives).</p>

Opportunities

- Could become the **de facto standard** that even incumbents adopt (e.g. integrate protocol into Cloudflare or browsers for broad reach).
- Leverage **regulatory endorsement**: perhaps become an official compliance mechanism under AI Act, driving adoption by policy.
- Use tokenization for innovative models (e.g. **Crawl-NFTs** for rights ⁵⁵, or secondary markets for data licenses).
- Expand beyond text: the standard could extend to image, audio content usage by AI, opening new verticals.

- Monetize its huge network in a new way; **increase ARPU** of existing clients by adding this value-add service.
- Possibly become the “App Store” for data: taking a cut of every AI content transaction (diversifying from pure CDN fees).
- If succeeds, it locks in customers further (stickiness) – publishers less likely to leave Cloudflare if it’s a revenue source.
- Can partner with large publishers early to set industry norms (as seen with those who signed on to block by default) ⁵⁶ ⁵⁷.

- Can entrench itself as the **main marketplace** if it continues to scale both sides (network effects).
- Potential to offer **analytics and insights** from AI traffic that publishers will pay for (beyond just tolls, becoming a data provider).
- Could move up the value chain: e.g. facilitate full licensing deals or subscription bundles for AI access, not just per-call – becoming an exchange of record.
- Acquisition potential: could be bought by a larger platform (potential lucrative exit) to instantly give them a foothold in this emerging market ⁵⁸.

- Turn a threat into PR win: by voluntarily paying publishers, Google could improve relationships and preempt regulation.
- **End-to-end control**: opportunity to integrate paid access directly into search results or browser (e.g. Chrome could handle 402 payments seamlessly), setting its own standard.
- If Google charges its AI competitors for access to YouTube/Maps/ etc data, it can even create a **revenue stream from others’ AI** usage.
- Could leverage YouTube’s content-ID and monetization expertise, applying a similar model for text on the web (since Google has experience managing a

Factor	Startup (Open Pay-Per-Crawl Protocol)	Cloudflare (Pay Per Crawl)	TollBit (AI toll marketplace)	Google (Potential Entrant)
				creator ecosystem with payments).

Threats

- Incumbent retaliation:

Cloudflare or browsers could refuse to support the protocol's signals (e.g. ignoring 402 responses or signatures) which would stymie adoption.
-

Fragmentation: If multiple standards proliferate (Cloudflare's, TollBit's, etc.), the startup's solution might never achieve critical adoption – AI players may cherry-pick the easiest/cheapest route.
-

Regulatory shifts: If laws impose a specific solution or collective licensing mandate, a startup-driven protocol might be sidelined by mandated frameworks (or big consortia).
-

Security/abuse: A breach or exploit in the payment system (e.g. stolen keys, fraud) early on could scare off would-be users and give the concept a bad rap.

- Competition from other

infrastructure: If Fastly/Akamai join forces with an open solution or a coalition, Cloudflare could lose its unique selling point.
- Could face **pushback from Big Tech** clients who don't want Cloudflare dictating terms or costs for crawling (e.g. Google might route around Cloudflare's network).
-

Margin considerations: handling micropayments and disbursements might be outside Cloudflare's core competency; any technical failure could harm its main reputation (security/perf).
- Regulators might view Cloudflare's control over such a marketplace as too much centralized power over information access.

- Low adoption

risk: AI firms might find ways around TollBit's detection (e.g. disguising as humans), reducing its value to publishers.
-

Cloudflare's free alternative could undercut TollBit's pitch; if Cloudflare offers this to sites at minimal cost, TollBit must justify its fees and independence.
- Legal outcomes could make TollBit's contracts unenforceable or unnecessary (e.g. if a precedent allows reasonable scraping, publishers might just block or open up without paying toll).
-

Scalability of business: if revenue per publisher stays low (as some fear ²⁰), TollBit might not sustain its marketplace economics and burn through cash.

- Unified publisher

pushback: If the industry perceives Google's solution as self-serving, publishers could collectively refuse and stick with independent methods (as they did when negotiating news licensing).
-

Regulatory: Google is under antitrust and could be constrained from leveraging its dominance to impose a proprietary system here (especially in EU).
- If Google's own AI efforts falter or policy forces separation of search and AI, its incentive to solve this might wane.
- A misstep (like underpaying creators or using this to scrape more) could lead to worse PR and renewed calls for strict

Factor	Startup (Open Pay-Per-Crawl Protocol)	Cloudflare (Pay Per Crawl)	TollBit (AI toll marketplace)	Google (Potential Entrant)
				regulation or publisher boycotts.

Summary: The startup's main **competitive differentiator** is its vision of an **open, blockchain-based standard** that isn't beholden to any single gatekeeper ⁵⁹. This could be a true moat if it gains momentum – neither Cloudflare nor TollBit can fully decentralize without hurting their models (Cloudflare would lose lock-in; TollBit its intermediary role) ⁶⁰. However, those incumbents have a head-start in market presence. Cloudflare in particular could capture a huge user base quickly, while TollBit has been tailoring the solution and gathering data for longer. Major tech companies loom, but are likely to react rather than proactively enter at this stage (except by funding or quietly partnering, as Google's Jeff Dean did with TollBit ⁴⁴). The competitive landscape is thus **fluid and rapidly validating** the need for a solution. The protocol must prove its approach before incumbents either co-opt it or sew up the market with their own offerings.

4. Customer Validation

The core customer segments are **web content publishers** (supply side) and **AI developers/vendors** (demand side). We examine evidence of their willingness to participate in a pay-per-crawl system:

Publisher Willingness to Pay/Participate: There is strong evidence that publishers are not only willing but **eager to demand payment** for AI access: - **Behavioral evidence:** As noted, 73% of publishers have actively blocked AI bots ¹³ – a drastic step indicating they'd rather lose marginal traffic than be harvested for free. This suggests a readiness to enforce paywalls if a viable mechanism exists. Further, dozens of big publishers (Conde Nast, AP, *Fortune*, *Adweek*, etc.) have signed up to block AI by default via Cloudflare, explicitly supporting a "permission-based approach to crawling" ¹⁶ ⁶¹. These are strong signals of latent demand for a solution that can **monetize or control AI access**. - **Deals and pilots:** Some publishers have proactively struck deals which implicitly put a **price on their data**. OpenAI's deals with *The Atlantic* and *Vox Media* (and reportedly *News Corp*) show publishers will engage if compensation is offered ¹⁷ ¹⁸. Those who aren't getting deals are exploring platforms like TollBit – **2,000+ sites** signing on indicates significant interest in any path to monetize bots ²¹. While actual revenue figures are not public, the enrollment itself is a form of validation. - **Publisher statements:** Industry leaders have been vocal. For example, a publishing exec told *Digiday* that **charging RAG bots for scraping will be key to future revenue growth** (even if it's early) ⁶². The Financial Times' submission to UK's House of Lords condemned the status quo of "simulacrams" being created from publisher content without remuneration ⁷. These sentiments underscore that publishers **want to get paid** by AI users much like they do by aggregators or syndication today.

The key question is not desire – that's clear – but **ability to pay vs. willingness to accept trade-offs**. Smaller publishers may lack tech resources to implement a new system (only ~22% of SMB sites could manage custom fee structures, according to Forrester ⁶³). Also, some publishers worry about losing any remaining Google traffic if they aggressively gate content (they fear "hiding" from good bots while blocking bad ones ⁶⁴ ⁶⁵). Early adopters will likely be those *most directly hit* by AI scraping (news, information databases) and with resources to experiment (national newspapers, large digital media networks, etc.).

Likely Early Adopter Persona: A prototypical early adopter would be a “**Digitally savvy mid-to-large publisher**” – for instance, the CTO or Head of Audience at a major news organization or content network. This persona: - Already **observes significant bot traffic** in analytics and has taken some steps (blocking or legal). - Feels urgent revenue pressure from declining referral traffic and is actively looking for new monetization streams. - Has a dedicated tech team or uses a CDN where integration friction is low (so they can trial a new protocol or service easily). - Organizationally, has support from leadership (e.g. CEO/COO of the media company) to “make tech giants pay” – many CEOs of publishing companies have publicly called for solutions, indicating top-down buy-in.

Examples might include digital-forward outlets like *Insider*, *NY Times*, *Thomson Reuters*, or large magazine conglomerates – essentially those who have the scale of content that AI wants, and the clout to be early movers. Another early adopter category could be **platforms with user-generated content** that is valuable (Reddit, StackOverflow have already attempted to charge for API access in lieu of scraping). They are not traditional publishers but are highly motivated to monetize AI usage of their data.

AI Vendor Willingness to Pay: On the demand side, willingness is more begrudging, but there are signals: - **Voluntary payments so far:** A few AI companies have shown willingness by inking licensing deals (OpenAI, Microsoft via Bing, and smaller ones like DuckDuckGo or You.com which licensed content for their answers). OpenAI in particular is paying multiple publishers and agencies (AP, etc.) ¹⁷, indicating a recognition that *paying for reliable data* can be worth it – whether for legal cover or data quality. - **Response to blocking:** When faced with blocked access, some AI startups have chosen to negotiate. For example, the startup **Perplexity.ai** has reportedly established revenue-share deals with top-tier sites via TollBit ⁶⁶. This suggests at least niche AI players will pay if it's the only way to get high-value content (especially domain-specific or real-time info). - **Industry direction:** As the AI field matures, there's an understanding that “raw web data isn't free forever.” Anthropic's CEO and others have hinted they are open to frameworks that compensate creators, especially under regulatory pressure. And many smaller AI firms (who cannot afford lawsuits) see a marketplace as an **easier route than negotiating hundreds of licenses individually** ⁶⁷. TollBit's founder noted AI companies aren't unwilling to pay, they just lack a scalable way to do so now ⁶⁸.

That said, **most AI vendors will not pay until forced** – their current behavior is to scrape what they can until blocked or sued ⁶⁹ ⁷⁰. In all likelihood, early willing participants on the AI side will be: - **Startup model providers** who need specific data (and can't afford legal risks) – they may join a marketplace to get, say, access to *Time.com* archives or StackOverflow content in a compliant way rather than going against a block. - **APIs and aggregators** that already position themselves as ethical or high-quality data sources (some smaller search engines, etc., might pay to differentiate on having licensed content). - The big players (OpenAI, Google, Meta) will resist until there's enough publisher coordination or regulatory mandate to push them to the table. They've shown willingness to pay only in targeted ways so far.

Customer Discovery Red Flags: To gauge true interest (versus lip service) from potential customers, we would pose discovery questions. Three **red-flag questions** to test a publisher's genuine interest might be:

1. **“Are you actively tracking AI bot traffic and its impact on your site's revenue?”** – If a publisher hasn't even quantified the problem (no idea how many GPTBot hits or what their referral drop is), it's a red flag. It suggests low awareness or urgency, meaning they might not be ready to implement a solution. An engaged answer (e.g. “Yes, we saw a 5% traffic drop and X million bot hits last quarter”) indicates pain awareness and likely openness to a remedy. A nonchalant “Not really” would signal a tougher sell.

2. **“What would you do if an AI crawler offered you \$0.01 per article view? Would you allow it, or worry about other consequences?”** – This probes their willingness to trade open access for money. If the response is “Yes, absolutely, we’d take any compensation” – it’s a green light that monetization is enticing. But if they say “We’re concerned that might hurt our Google rankings or user experience,” that’s a red flag: they might view participation as too risky or not worth upsetting the apple cart. We need publishers who are ready to experiment, not ones highly hesitant to charge anyone for access.
3. **“Have you allocated technical or legal resources in the next 6 months to address AI crawling (e.g. updating terms, adding paywall tech, pursuing agreements)?”** – This checks if the problem is high on their roadmap. A publisher who says “Yes, our legal team is drafting new terms and we’re evaluating tech solutions (or we joined an alliance)” is signaling strong intent – they are actively seeking a fix, so they’d likely pilot a protocol. If they answer “No, we have no specific plans yet,” it implies they aren’t ready to act (red flag that they might just be in wait-and-see mode).

Similarly, for AI vendors, a discovery question could be: “What’s your plan if the top 100 news sites block your crawler tomorrow?” – if they shrug it off (red flag: they’ll try to evade or use alternative data), versus “we’d need a solution, we’re open to paying if we must” (sign of eventual willingness). Overall, these questions gauge whether the prospective users perceive the pain acutely enough to change behavior (integrate a new protocol or pay money) – without that, any expressed interest might be mere window dressing.

5. Technology & Moat

The Pay-Per-Crawl protocol relies on several **critical technical assumptions** that must hold true for it to succeed:

- **Micropayments at Scale are Feasible and Cheap:** The model assumes that on-chain or blockchain-based micropayments can be executed in massive volumes (potentially millions of requests per day) with negligible fees and latency. The analysis projects using Layer-2 networks (like Base or similar) with fees **<\$0.0001 per transaction** ⁵¹ ⁵², which is incredibly low. This needs to be proven in practice – any blockchain congestion or fee surge could wreck the economics. Additionally, the payment process must not introduce noticeable delay in serving content (sub-100ms ideally). The **x402 standard** (HTTP 402 + blockchain payment in headers) is unproven at large scale; it’s critical that it adds minimal overhead and that off-chain components (e.g. oracle or wallet services) don’t bottleneck. This assumption appears technically plausible given modern L2 throughput, but only a live test at scale will validate it.
- **Secure Crawler Identification:** The protocol banks on **Decentralized Identifiers (DIDs)** and cryptographic signatures to ensure the entity hitting a publisher’s URL is genuinely an approved, paying AI crawler and not a spoof ⁷¹ ⁷². This is vital to prevent trivial workarounds (e.g. a bot pretending to be Googlebot to bypass pay). The use of Ed25519 keys and signed headers is a solid approach in theory ⁷³ ⁷⁴, but it assumes crawlers will implement this correctly and that publishers will reject any request without proper signature. It also assumes the DID registry won’t be compromised. Essentially, the security of the identity layer must be ironclad – a single high-profile spoofing incident (an AI bot faking another’s ID to get free content) could undermine trust in the system. The protocol’s moat relies on this being **bulletproof and easier than trying to evade**.

- **Automated Contract Enforcement (Smart Licenses):** The idea of a **tokenized license (NFT) attached to each crawl transaction** is to provide a clear, enforceable record that content was accessed under certain terms ⁵⁵. This needs to satisfy both technological and legal criteria. Technically, each transaction must log the URL (or a hash of it) on-chain with the terms. This raises concerns about **privacy and scale** (logging every URL could be heavy, even as hashes). The assumption is that a tamper-proof public log can be maintained without exposing sensitive data and without exorbitant storage costs. It's a key part of the value prop ("auditable trail"), so the engineering of this logging/audit subsystem is critical. On the legal side, it assumes such on-chain records will indeed be recognized by courts as evidence of a contract. The framework draws on e-signature laws and is probably sound ⁵⁵, but until tested, it's an assumption that these tokenized contracts will hold up (a court could one day say "this isn't a valid contract for copyright license" – a lot rides on getting this right).
- **Integration Simplicity:** For publishers, a crucial technical assumption is that integrating the protocol is as easy as "deploying a simple edge worker" or similar ⁷⁵. The solution must not require complex software installs or deep blockchain knowledge on the publisher's part. If it's not plug-and-play (think a WordPress plugin or a one-line cloud function), adoption will stall. The founder memo assumes integration friction can be reduced to near-zero with good tooling ⁷⁵. Achieving that is essential – likely meaning providing client libraries, CDN integration, or a managed service layer on top of the raw protocol.
- **AI Side Implementation:** On the AI vendor side, the assumption is they can integrate **automated payments via smart wallets (ERC-4337)** and handle the 402 responses gracefully ⁷⁶. This means every compliant crawler needs a crypto wallet, needs to watch for 402 payment-required responses, and then programmatically execute a stablecoin payment in real-time to retrieve content. That's a non-trivial engineering ask. The hypothesis is that vendors will adopt DIDs and x402 once enough publishers demand it ⁷⁷, but technically, it requires inserting a new loop in the crawler's fetching logic. If that integration is seen as too complex or slow (especially for companies not previously using crypto), they might delay or use alternate methods. So the protocol's design must make it as easy as adding a module to their crawler libraries – possibly providing reference implementations to reduce this barrier.

In terms of **moat and defensibility**: much of the technology is built on **open standards** (HTTP, blockchain protocols) which in theory anyone can replicate. The startup's strategy for moat is likely **network effect and first-mover advantage in standard-setting**. If they can get enough adoption, the standard itself becomes the moat (similar to how HTTP or TCP/IP are open but you gain power by being the one to implement and host services around them). The founder argues incumbents like Cloudflare or TollBit would have to "cannibalize" their models to copy a fully open approach ⁶⁰. Indeed, Cloudflare's system could technically integrate stablecoin payments and DIDs (nothing stops them from adding that within 6–12 months), but doing so in an open way (where Cloudflare isn't the intermediary for funds) might cut them out of a revenue loop. So **will incumbents replicate in <12 months?** They can replicate the *features* (charging bots) easily – in fact Cloudflare already has, and TollBit as well. But replicating the **open architecture** is another matter: it would mean, say, Cloudflare implementing x402 in a way that any wallet can pay and not requiring Cloudflare membership. That's against their short-term business interest. Thus, the moat is more strategic than purely technical: the startup is championing openness as a wedge. However, from a technology perspective, **nothing proprietary is truly unbuildable by a well-funded team** – identity standards, micropayments, audit logs are all in the public domain or easily accessible libraries. A giant like Amazon or

Microsoft could stand up a similar crypto-based payment system relatively quickly if they saw benefit. This means the startup's advantage must come from being **first to achieve critical mass** and possibly protecting some specific know-how (they might have some patentable components, though none obvious yet beyond maybe the specific use of NFT licensing).

Speaking of **IP and patents**: We should examine if there are existing patents or conflicts. Charging for bot access and using HTTP 402 isn't entirely new – 402 has been a reserved code for "Payment Required" for decades, and companies have envisioned micropayment schemes in the past (though not widely implemented). Cloudflare's approach is presumably not patented (they've made it sound like an open experiment, and 402 is an IETF standard code). TollBit might have filed patents on aspects of their platform (e.g. bot identification algorithms or the paywall redirect mechanism), but since TollBit operates more as a service, their innovation is likely not heavily patented or, if so, could possibly be designed around by using different technical means. The protocol's use of specific standards (DID, HTTP signatures) leverages public specs – that's good for freedom to operate. There could be **patents around web monetization** – for example, Google holds many patents on web advertising and maybe content licensing tech. But licensing content via contract is such a fundamental action that it's hard to fully patent. One area to watch: **digital rights management and licensing standards**. If there's any existing standard (perhaps something by W3C or others) on automated content licensing, any overlap should be checked. As of now, we're not aware of a directly conflicting patent. The startup might consider filing patents on their specific method of tying smart contracts to HTTP requests, to prevent copycats from literally cloning their implementation. But given the ethos of openness, they might rely on being a standard-setter rather than IP locks.

In summary, the technology is **promising but unproven at scale**. The moat will not come from secret algorithms but from successfully building an **ecosystem** around these protocols. If they prove out easy integration and show real volume on-chain with negligible cost, that itself becomes a barrier (others will then have to catch up and convince users to switch). There is a **first-mover advantage in establishing the de facto protocol** – much like how OAuth or SSL became standards partly by who rolled them out. The risk is if an incumbent or consortium quickly co-opts the idea and publishers follow that before this startup's solution gets traction.

6. Go-to-Market & Economics

Achieving adoption of a two-sided marketplace requires savvy go-to-market (GTM) strategy and careful unit economics. We draw parallels from analogous markets (web infrastructure, ad-tech, and Web3 tools) to chart a viable plan:

Acquisition Channels:

- **Enterprise Sales to Flagship Publishers:** In the early phase, manual outreach to high-value publishers will be key. This resembles enterprise SaaS sales. Likely channels include industry conferences (e.g. Digital Content Next, IAB events where publishers discuss AI), direct introductions via the News/Media Alliance, and leveraging investor networks to get warm intros to major media CTOs. Benchmark: In the advertising tech world, acquiring a large publisher often takes 3–6 months sales cycle and tens of thousands in CAC (travel, custom integration support). We might expect similar here, though the pitch is revenue generation (a plus). The **CAC** to land a *NYTimes* or *CNN* might be high, but their adoption would be game-changing for credibility.

- **Integration Partnerships (CDNs, CMS, security firms):** A leveraged channel is partnering with services publishers already use. The Fastly-TollBit partnership ⁴⁵ is a great example: it allows fast deployment through an existing vendor. Similarly, working with **Cloudflare's open platform** (if they allow third-party apps or workers) or with CMS platforms like WordPress VIP, Drupal, etc., can embed the protocol option directly in workflows. Analogy: in web infrastructure, companies like Cloudflare and Shopify often have marketplaces for add-ons – being present there can drastically lower CAC by piggybacking on another's distribution. If our protocol can offer a plugin on **popular web hosting or CDN dashboards**, a publisher can sign up in a click. This is likely lower CAC per customer (nearly self-serve if done right), though it might involve revenue-sharing with the partner.
- **Developer/Community Adoption (Web3 angle):** On the AI vendor side, appealing to developers is crucial. Open-sourcing parts of the stack or providing a robust API/SDK can drive adoption by AI startups. Posting libraries on GitHub, doing demos at AI dev meetups, and engaging with the Web3 developer community (which is enthusiastic about decentralized web) could attract early technical users who implement the client side for us. This strategy has worked for some Web3 protocols (e.g. Livepeer in decentralized video streaming gained dev traction through hackathons and grants). While AI giants won't join via a hackathon, the smaller players might, and they can be the initial demand side participants.
- **Thought Leadership & Policy Channels:** Given the regulatory push, part of GTM is influencing the narrative. Publishing whitepapers or contributing to policy discussions can position the protocol as the **go-to compliance solution**. If we can get mentioned in an EU AI Act guidance or by a trade body (IAB, etc.) as a solution, that's a huge acquisition driver (not in a traditional paid CAC sense, but as a trust signal). For example, if the EU were to reference using auditable data marketplaces to fulfill Article 10 data governance ⁷⁸ ⁷⁹, companies will seek solutions like this proactively.

CAC Benchmarks: As an early infrastructure product, CAC will likely be **lumpy**: - For **large publishers (enterprise)**: CAC could be \$10k-\$50k each when factoring sales engineer time, legal review cycles, possibly even incentives (e.g. offering a revenue floor guarantee to convince them to onboard). This is analogous to an enterprise cloud service sale. However, each such publisher might bring tens or hundreds of thousands in annual revenue once the system is live (thus justified). - For **smaller publishers (self-serve)**: The aim is to drive CAC down by offering easy sign-up and perhaps a freemium model (e.g. free monitoring of AI traffic, pay or share rev when you start charging bots). Many ad tech companies use this approach: provide a free analytics tool to get publishers in the door (TollBit itself offers a free "Content Cache" service up to a point ⁸⁰). CAC for self-serve can be just marketing spend – perhaps a few hundred dollars in digital campaigns or webinars to attract a mid-tier publisher, which is minor if LTV pans out. - On the **AI vendor side**, CAC might be more about integration effort than marketing dollars. Convincing AI companies to join will involve technical support, maybe custom features (if a vendor needs usage caps or special reporting). It's somewhat analogous to supply-side platform hooking up with demand-side in ad exchanges – often it's relationship-driven rather than big marketing spend. Early on, CAC here might just be founder time and integration engineering (cost of salaries).

Revenue Model & Unit Economics: Likely the protocol (or the startup operating it) will take a **transaction fee** on each crawl payment – say 5-10%. We can model unit economics with some assumptions: - **Average price per crawl:** This could vary widely (\$0.001 to \$0.01 perhaps). Let's assume an average of **\$0.005** per request as a midpoint (some pages high-value at 1¢, others fractions of a cent). - **Volume per publisher:** A medium publisher might get 1 million AI crawl hits per month; a top news site could see 10+ million. Using

1 million as a rough average for those onboarded (to be conservative initially). - **Gross Revenue per publisher:** 1,000,000 requests * \$0.005 = **\$5,000/month** gross paid by AI crawlers to that publisher. For a larger one with 10M requests, \$50k/month. - **Take rate:** At 10%, the platform earns \$500/month from the mid-tier publisher, \$5k from the larger in these examples.

If the startup onboarded 100 such publishers (mix of sizes), the **monthly gross merchandise value (GMV)** might be on the order of \$500k (100 * \$5k average), and the startup's revenue at 10% = **\$50k/month** (\$600k annual). That's a starting point. As more publishers and higher volumes come in, this scales linearly. The key to improving unit economics is **driving up volume and transaction frequency** (the marginal cost per transaction is tiny if on-chain fees are minimal, so scaling doesn't hurt gross margin).

Speaking of **gross margin**: This should be very high – likely **85-90%+** – because the cost of facilitating a transaction is just some blockchain gas (which is paid by the transacting parties in many designs) and overhead of running the platform's off-chain components (matchmaking, dashboards). If the startup covers some gas or runs nodes, that's maybe a few percent of the fee. The heavy investment is in R&D and support initially, not cost of goods, so gross margins resemble software/network business.

LTV and Payback: If a publisher yields \$500/month to the startup, that's \$6k/year. LTV (assuming a decent retention, say 3-5 years lifetime) could be \$18k–\$30k. If CAC per major publisher was \$10k, that's a 1.5x to 3x LTV/CAC, which is acceptable but needs to grow (SaaS benchmarks aim for >3x LTV/CAC). However, if the volume ramps up as the AI usage grows (which it likely will, perhaps doubling each year for a while), the revenue per publisher could increase without additional CAC, dramatically improving LTV. Also, once integrated, switching costs are moderate (publishers could switch to a competitor, but if they have smart contracts and on-chain records tied up, and they see revenue flowing, they'll be reluctant to churn without cause).

Sensitivity: The economics are sensitive to: - **Price per crawl:** If the average price is only \$0.001 instead of \$0.005 (maybe due to competitive pricing wars or publishers setting low fees to encourage usage), then revenues are 5x lower. The model might need a high volume of crawls to compensate or a higher take rate (which could be hard to justify). - **Adoption ramp:** If only half the anticipated volume onboarded in the first 2 years, revenue stays low while fixed costs (team, platform dev) accumulate, potentially requiring more capital. - **Take rate changes:** If a competitor like Cloudflare takes 0% fee as a loss leader (they might, to attract users onto their core platform) ³², the startup might be pressured to lower its take from 10% to, say, 5% or 2%. That squeezes margins significantly unless total volume is enormous.

Sales Cycle & Friction: For publishers, expect some **friction in the sales cycle**: - **Legal review:** Each big publisher will have lawyers comb through how this protocol works (Is it safe? Does it conflict with our terms? Are we liable if something goes wrong?). The novelty (smart contracts, crypto) could elongate this process. Educating legal teams is part of the "sales" effort. - **Technical integration cycle:** Even if integration is "one-click", enterprises often do staging tests, security audits, etc. That could add weeks or months. Some might wait to see a peer implement it first (nobody wants to be the very first in case something breaks). - **Internal alignment:** At a media company, the CTO, Chief Digital Officer, and maybe editorial leadership all need to agree to this approach (especially if it involves potentially blocking certain bots which could have PR implications). The sales process needs to address concerns from multiple stakeholders.

These frictions mean a slower ramp – a realistic goal might be landing a handful of marquee publishers in year 1 as proof points, rather than hundreds.

Churn Risk: Churn on the publisher side will depend on delivered value: - If the revenue trickles in too slowly or remains small, publishers might lose interest after the “pilot fever” wears off. For instance, if a site integrates and only makes \$50 in the first month, they may decide it’s not worth the complexity and disable it (especially if they didn’t block bots fully). - **Competition-driven churn:** If an incumbent offers a sweeter deal (e.g. Cloudflare waives all fees or bundles the service free with CDN), a publisher might churn off the startup’s platform in favor of that. - **External factors:** If legal outcomes suddenly guarantee them payment via another channel (say a collective licensing law gets passed), they might exit the marketplace in favor of that mandate.

On the AI side, churn is about continued participation: - An AI company might integrate initially (to be a good actor or as a pilot), but if they find they’re spending too much on crawl fees, they might try to circumvent or exit the system. We should expect some to test the waters then pull back if ROI isn’t clear for them.

To mitigate churn, the GTM has to **manage expectations** (this is a new revenue stream that may start small but grow) and ideally lock in customers via success (once a publisher sees meaningful monthly income and relies on it, they’ll stay). Providing ancillary value like **analytics** (“know who’s crawling you and how often”) even before revenue scales can keep publishers engaged – they get insight even if payments are modest at first.

In terms of **analogy**, this is similar to early ad networks or affiliate programs: at first they might only earn a site a few bucks, but the data and potential kept them trying until it scaled up. Ensuring a good publisher experience (timely payouts, transparent reporting, low effort) is key to reducing churn risk.

7. Execution Risk Map

Launching this protocol is a complex endeavor requiring a mix of technical, legal, and market-building skills. Here we identify key execution risks, including team capability gaps, timeline challenges, and capital needs:

Team & Talent Gaps: The founding team must cover an unusually broad domain (crypto engineering, web standards, B2B sales to publishers, and legal/contracts). Likely gaps: - **Enterprise Sales & Publisher Relations:** If the team is heavy on blockchain developers and light on media industry experience, that’s a gap. Selling to publishers and navigating their concerns requires domain expertise (understanding CMS platforms, ad ops, etc.). Hiring a bizdev lead or advisor with deep publisher network (ex-publishing executive) will be crucial. - **Legal/Regulatory Expertise:** Given the legal novelty, having an in-house counsel or access to top legal minds (IP, contract law, data privacy) is needed. If currently the legal strategy is just theoretical, that’s a risk. Bringing on a legal advisor familiar with E-SIGN, AI Act, and case law would help de-risk the contract enforceability aspect. - **Scalability & Security Engineering:** The blockchain and payment side might be handled by a small dev team now. But as this grows, **smart contract security** and high-scale infrastructure become critical. Any talent gap in devops or security auditing could lead to vulnerabilities. Filling out the engineering team with specialists (e.g. a security engineer to rigorously test the payment system, a scalability engineer for the API gateway) is important. - **Product UX for Integration:** There’s a risk if the team lacks a good product designer or UX engineer to make integration seamless. If it’s too geeky or rough, adoption suffers. Ensuring the team includes someone who can build polished tools (maybe a developer relations person to create clear docs, SDKs) will address this.

Milestones & Timeline: - *0-6 months: Prototype/MVP and Pilot* – Goal: Build a working payment flow (maybe on a testnet or L2) where a sample publisher can gate content and a sample crawler pays and accesses. Also, sign at least 1-2 pilot publishers and one AI startup to test in a controlled environment. **Blockers:** Getting pilots to commit (they'll want to see it working first – catch-22). Also, regulatory uncertainty might make some publishers hesitant to pilot without their legal's blessing. We might need to offer strong assurances or indemnities to early pilots to get them on board. - *6-12 months: "Atomic Network" launch* – Focus on a niche, e.g., a group of tech news sites plus a couple of AI research companies. Show that within that micro-network, the loop works (publishers earned \$\$, AI got data conveniently). Milestone could be something like: 50 publishers, 5 AI consumers actively transacting. **Blockers:** Cold start – convincing enough publishers to block non-paying bots in that niche simultaneously. It may require coordinating a sort of mini "strike" where they all agree to enforce payment, creating incentive for the AI side. Herding those cats on timing and terms is tough. Another potential blocker: technical scaling issues (if, say, gas fees spike or an exploit is found when more users join, it could delay progress). - *12-24 months: Scale to broader market or verticals* – Aim to expand beyond the initial niche to either other content verticals or more geographies. Perhaps this coincides with regulatory milestones (e.g., AI Act enforcement – we'd want to be ready to capture EU enterprise interest by that time). Milestones here include closing a partnership with a major platform (maybe a CDN or a browser company) for distribution, and reaching a sustainable GMV that attracts Series A/B funding. **Blockers:** By this stage, competition might react – Cloudflare could launch general availability of their service, undercutting our narrative. Also, legal challenges could arise: e.g., an AI company might proactively sue to challenge the enforceability of such contracts. Dealing with that while scaling would be a huge distraction. There's also the execution risk of expanding the team and operations too quickly if interest surges (startups can stumble if they scale org faster than product-market fit, leading to burn without matching revenue).

We should maintain a **risk register** mapping these with contingencies: - **Talent risk mitigation:** Plan to hire or contract for key expertise (like bring in a publishing industry advisor by Q1, hire a lead engineer from a top crypto project by Q2, etc.). - **Technology risk mitigation:** Do thorough audits and maybe limit initial financial exposure (cap payments or use test currency) until security is proven. Also have a fallback if L2 fails – e.g., be ready to switch to another chain or even a centralized ledger temporarily. - **Marketplace liquidity risk:** If one side adoption lags, consider interim measures (like a **publisher subsidy or a guaranteed minimum payout** funded by us to encourage them to join, essentially seeding the marketplace).

Capital Intensity & Funding Plan: This venture has a significant **market-building cost**. It's not pure software where you build and they come – we must invest in integrations, support, and evangelism. That said, compared to something like a hardware startup, the capital needs are moderate. The largest expenses will be: - **Personnel:** Engineering, sales, compliance – high-quality talent in blockchain and in media sales is expensive. We're likely looking at Silicon Valley-level salaries for engineers and New York-level for media partnerships. A team of, say, 15-20 by next year could burn a few million a year easily. - **Legal & Lobbying:** We might need to spend on legal opinions, possibly lobby efforts or membership in alliances to shape standards. That can add a few hundred thousand. - **Infrastructure:** Running nodes, data storage for audit logs, etc. – initially minor, but if we log millions of transactions, costs for reliable infrastructure (nodes, cloud servers, monitoring) will grow. Still, these are in the tens of thousands range, not millions, at early scale.

To be **default alive** (profitable or at least break-even on current funding) is likely not feasible until a critical mass of transactions occur, which could be 2-3 years out. More realistically, this will need to raise a **Series A** once the concept is proven in pilot form to fund the scale-up phase. Funding milestones might be: - **Seed**

(current): Build MVP, get first pilots = prove technical feasibility and initial demand. (Likely already raised or in progress). - **Series A (~12-18 months in):** Raise to accelerate adoption – funds for hiring sales, expanding network, possibly marketing to publishers. Achieved if we can show some recurring revenue or at least significant usage in a controlled environment. At Series A, investors will look for signs of the flywheel starting (maybe some GMV and growth rate). - **Series B and beyond:** If growth takes off, more funding might go toward international expansion (bringing in EU publishers ahead of AI Act, etc.), and defending against competitors (could mean spending more on R&D, or even acquisitions – e.g., acquiring a smaller competitor or key tech).

Capital Intensity vs ROI: The business can become **capital efficient** once the marketplace is liquid – since each new transaction costs little to service, margins are high. But getting there requires front-loaded investment in technology and network building (which is why outside funding is needed). We should plan scenario burn rates to ensure we don't run out of cash before the network effect kicks in. Being "default alive" prematurely might mean not investing enough in growth to beat competitors. More likely, we operate at a **loss through Series A/B** while building the network, aiming for a **break-even after significant scale** (perhaps by Series B when annual transactions might be in the hundreds of millions, generating sustainable revenue).

In short, execution will involve carefully balancing speed (to outrun competitors and seize mindshare) with prudence (ensuring legal and technical robustness). Missing a key hire or underestimating a regulatory hurdle could derail timelines. Conversely, hitting each milestone (MVP > niche network > broader adoption) on schedule will likely require raising successive funding, which demands continually proving out assumptions to investors.

8. Scenario Stress Tests

It's important to envision how this venture might play out under different scenarios. We outline three outcomes – Best case, Base case, and Downside – and how the team could navigate each:

Best Case (95th Percentile – “Breakout Success”):

In this scenario, the protocol achieves a *flywheel effect* early and dominates its niche: - **Year 1:** The startup secures a handful of marquee publishers (say 5 of the top 20 news sites) who **simultaneously implement pay-per-crawl** on their content. This coordinated move, perhaps aided by a public stance from a publishers' association, forces at least one major AI player (e.g. OpenAI or a consortium of smaller AI startups) to participate to avoid losing critical data. The MVP works with only minor hiccups: AI crawlers successfully integrate the payment API, and stablecoin transactions flow smoothly at scale (demonstrating that \$0.0001 cost indeed). By the end of year 1, the protocol has processed, say, **50 million paid requests**, proving technical viability and netting publishers meaningful revenue. - **Year 2:** Building on that credibility, adoption **accelerates**. Regulatory tailwinds kick in – for instance, as soon as the EU AI Act is official, European publishers flock to the solution as a ready compliance measure. The startup is able to announce partnerships with a major CDN (perhaps Fastly or Cloudflare grudgingly supports the standard due to client demand) and an integration with a popular web CMS. This makes onboarding the next 1000 publishers much faster (self-serve onboarding for mid-size sites becomes reality). On the AI side, even Google begins to experiment with the protocol in Europe to comply with regulators (maybe Google uses it for a subset of content or as a PR gesture). By the end of year 2, the marketplace has **thousands of publishers and dozens of AI clients**, with transaction volumes growing 20%+ month-over-month. - **KPI outcomes:** At this point, the company might be facilitating e.g. **\$100M in annual transactions**, shifting millions of dollars to

publishers. The startup's cut (if 5-10%) yields \$5-10M revenue, and the business approaches breakeven or profitability much faster than anticipated. Publishers start reporting that "AI licensing" is now, say, 5-10% of their digital revenue – a real new line item. - **Upside tactics:** To get here, the team likely executed near-flawlessly: leveraging one success to get the next (public case studies of early pilots converting skeptics), and perhaps raising a strong Series A to outpace competitors. They may have also engaged in savvy diplomacy – e.g., open-sourcing parts of the protocol to avoid fights and instead having would-be competitors join them (Cloudflare deciding to support the standard rather than fight it, turning into a collaborator). **Outcome:** The startup becomes the *default standard* for AI content access, akin to how Robots.txt became standard for crawling – except now it's Robots402 with payments. An exit or massive growth round could follow, but importantly the mission succeeds in reshaping the web's economic model in the AI era.

Base Case (50th Percentile – "Moderate Traction, Prove-Out"):

In the base case, the startup makes solid progress but also encounters headwinds that temper the speed of growth: - **Adoption:** A few high-profile publishers sign on, but perhaps not coordinated. For example, one major newspaper group and a couple of niche data providers start using the protocol. They do see some revenue (maybe tens of thousands of dollars in the first year – meaningful but not game-changing for them). AI companies, however, largely *resist*. Smaller AI startups join the marketplace, but the big ones continue to scrape via workarounds, so the volume of paid transactions is modest at first. This is similar to how some content paywall companies saw initial uptake: slow and one segment at a time. - **Traction:** By the end of year 1, maybe **200 publishers** (mostly small-mid ones and a few big names) have installed the code, but only a dozen AI bots are paying, mostly for niche or pilot use. The startup demonstrates the concept works – e.g., *a data science AI startup happily paid \$10k to access a bundle of tech blog content via the protocol instead of negotiating licenses* – but the overall market is still in early adopter phase. Regulatory changes loom but are not yet enforced, so many players take a "wait and see" stance. - **Financials:** The startup might be facilitating on the order of **\$5M in transactions annualized** by year 2, yielding maybe ~\$300-500k revenue to itself (assuming ~5-10% fee). This is respectable for a new platform but not explosive hockey-stick growth. They likely raised a Series A on the promise and need to raise a Series B to continue. - **Challenges:** Execution-wise, the team likely hit some snags – perhaps technical scaling issues that needed patching (maybe an incident where the blockchain network had an outage causing some payment failures, which they corrected with redundancies). Or resistance from AI firms required the startup to develop new features (like a **"data escrow" or capped spending features** to entice AI participation). Competition is present: Cloudflare's marketplace might be out of beta and wooing some of the same publishers, causing confusion and requiring the startup to differentiate by emphasizing decentralization and better economics. - **Mitigation:** In this scenario, the startup might pivot slightly or narrow focus to get a stronger beachhead. For instance, if broad web adoption is slow, they might focus on a high-value vertical like **financial news sites and fintech AI** – where they can show a strong case (since that data is very valuable and likely to be paid for). This creates a success story to later generalize. They also engage in alliances – maybe joining the IAB working group to ensure their protocol can become the recommended approach, securing future pipeline. - **Outcome:** The base case is essentially **plausible traction with steady growth, but not yet runaway**. The company remains in a solid position to continue, possibly needing more capital to push through the inertia. Investors would view it as "on track, but prove more metrics." The model is not broken, but also not yet validated at scale. It might be in a **"wait for the dam to break"** mode – prepared for a catalyst like a big legal decision or law that suddenly drives many holdouts to adopt. Patience and continued execution are key here.

Downside Case (5th Percentile – “Failure / Stalled Out”):

In a credible worst-case scenario, the concept fails to gain sustainable traction and the startup hits a wall: - **Publisher Stall-Out:** Perhaps publishers talk a lot about wanting to get paid, but when it comes time to implement, very few actually follow through. Some might start integration but then back off, either due to internal fears (legal says “let’s not be the test case in court”) or because alternative approaches emerge. For instance, big publishers might decide to pursue litigation and government regulation (a slower but potentially larger payout for them) instead of fiddling with technology solutions. Small publishers might decide blocking is enough and they don’t bother monetizing. So the supply side never reaches critical mass – maybe only 50 small sites onboard, which isn’t enticing to AI crawlers. - **AI Defiance:** Meanwhile, AI companies could double-down on **evasion tactics**. We know already bots try to mimic human agents ⁶⁴. In a downside scenario, the majority of AI players choose to route around the protocol entirely: they find proxy IPs, fake user agents, use cached content, etc., to avoid paying. This becomes a cat-and-mouse game that the startup can’t effectively win without broad publisher backing. One could imagine an outcome where, say, OpenAI and others publicly state, “We respect robots.txt and will negotiate deals, but we won’t integrate on-the-fly payments” – effectively stalling momentum. If AI crawlers continue getting what they need (via either ignoring blocks or because blocks weren’t widely implemented), the value proposition of the protocol falls apart (no one is forced to pay). - **Financial Strain:** In this scenario, revenue through the platform is minimal – perhaps a few experimental transactions, not enough to sustain a business. The startup burns through its seed/Series A money building the tech and evangelizing, but uptake is too slow. By, say, 18-24 months in, they’re low on cash with metrics below what investors want for the next round. For example, they might have <\$50k annual revenue and no clear hockey stick growth – a classic valley of death. - **Credible Failure Modes:** One failure mode is “**no network effect**” – you build it, but not enough came. Another could be an external shock: maybe a court rules that existing copyright law doesn’t cover data mining as infringement (giving AI companies a victory to continue scraping without paying). Or perhaps a large tech company introduces an alternative approach that undermines the startup – e.g. Google or Microsoft launches an “AI Publisher Fund” that pays top publishers a stipend, satisfying them enough that they don’t join a risky new platform, thereby keeping the status quo mostly. - **Mitigation & Pivot:** If facing this downside trajectory, the startup would need to pivot or drastically alter strategy to salvage value. Possible pivots: - Focus on a **different customer**: maybe pivot from publishers to data owners in other fields (like scientific data or APIs) where the model might still work. - Or become a **service company** instead of platform: for instance, reposition as an analytics tool (help publishers detect AI bot activity and optionally charge via simpler means like API keys) – essentially de-emphasize the micropayment vision and solve the awareness and blocking problem (something publishers might pay for as a SaaS even if AI won’t pay them). - Another mitigation: align with a bigger player. In failure mode, one might seek acquisition by an incumbent (Cloudflare, or even a publisher consortium) on the premise of “you have the tech and know-how, let us integrate it on our terms.” This might salvage some of the mission. - **Outcome:** The downside ends with either the startup winding down (if no pivot or buyer is found) or becoming a minor feature elsewhere. The broader concept of pay-per-crawl might lie dormant until possibly the environment changes enough in the future. The failure would likely be due to **market timing and coordination problems** rather than the tech itself – a classic case of being too early or not managing to align the key players.

Each of these scenarios has distinct signals. We should monitor early indicators: e.g., if by next year we haven’t closed any big publisher deals (pointing to downside), or if we see regulatory mandates causing inbound interest (pointing to best case). The team must stay agile, ready to capitalize on a best-case tailwind or pivot in a downside.

9. Deal-Killer Questions (Checklist)

Before investing further, we run through a checklist of potential *deal-killers* – if any of these are a definitive “Yes,” the venture’s viability could be in jeopardy:

- **Single Point of Failure? – Yes (in part).** The protocol itself is decentralized, but a few single points remain. One is **reliance on a specific blockchain network** for payments; if that network fails or fees spike, the whole system could stall (though they could mitigate by being chain-agnostic). Another is the **need for collective publisher action** – if a critical mass of publishers doesn’t act, the model fails (that’s an adoption point of failure). There’s no single server or company whose outage kills it (if truly decentralized), but practically, the startup operating it could be a single point of failure at early stage (e.g. if the company shuts down before fully open-sourcing/governance, the protocol might die on the vine). So, while no one publisher or AI client failure kills it, the project is dependent on some central dependencies (tech and market) in its current go-to-market approach.
- **Legal or Platform Risk of Being Outlawed/Banned? – Possibly, but not directly outlawed.** The concept of charging for content access is legal – in fact, it’s a property right. However, **platform policies** or laws could indirectly hobble it. For instance, big browser makers or OS (controlled by companies like Google) could refuse to implement or honor HTTP 402, effectively sidelining the user-agent mechanism (though not outlawing, just not supporting). Also, if an antitrust or competition authority viewed a coalition of publishers setting prices via a protocol as price-fixing, that could be a legal concern (collusion argument). But since it’s an open marketplace, that risk is lower (each publisher sets its own price). We should consider if future **copyright exemptions** or government mandates come into play: e.g., if the US introduced a compulsory license for AI training data (with a government-set rate), it might “outlaw” independent marketplaces by superseding them. Not likely imminent, but a possibility in the evolving policy landscape. In summary, there’s **regulatory uncertainty** but no known law forbidding what we’re doing. The biggest legal risk is if courts *don’t uphold* the contracts – not outlawing us, but making our service ineffective (if a judge said “scraping isn’t illegal and clicking accept on a 402 doesn’t create a binding license,” that undercuts the business severely).
- **Commodity/Price War Risk (Commodity Exposure)? – Yes, to a degree.** Web content can be a commodity in the eyes of AI training – if one source is too expensive, an AI might find similar info elsewhere cheaper. This means there is **downward pressure on prices**. If multiple systems or publishers compete, they could undercut each other (“race to the bottom” for crawl fees). For example, if *Reuters* charges and *AP* doesn’t, AI will use *AP*. In the long run, content that’s very unique (e.g. *NYT* investigations) has pricing power, but commodity content (basic how-tos, generic news that many outlets cover) could see a price war where only the cheapest sources get crawled. This exposes the model to **margin compression** – the protocol might end up facilitating micro-payments measured in fractions of a cent, limiting revenue. Also, incumbents like Cloudflare could engage in a price war on fees (0% marketplace fee vs our 5-10%). If our business relies on taking a cut, a price war could zero that out. So yes, there is a real risk that both the data and the service become commoditized. The strategy to mitigate is differentiation (quality content pool, value-added services on top of raw data) and network effect (if we aggregate exclusive must-have content, we avoid pure commodity status). But investors should note: we’re not dealing with a proprietary tech moat, so pricing power may not firmly be on our side; we have to earn it by network scale.

(Each of these items should be monitored closely. None are an immediate kill-switch yet, but they highlight areas to build contingency plans: multi-chain support to avoid single blockchain risk, strong legal counsel on contract enforceability, strategies to prevent a race-to-bottom pricing by emphasizing value over volume.)

10. Next-Step Experiments (0–90 days)

Given the early stage, we propose three lean experiments to validate key assumptions in the next 3 months. Each is low-cost and designed to maximally inform our direction:

Experiment 1: Publisher Landing Campaign & LOI Test

Goal: Validate publisher interest and identify early adopters willing to pilot.

Design: Create a simple microsite or landing page explaining the Pay-Per-Crawl solution, with a call-to-action for publishers (“Join the Beta: Get paid when AI crawls your site”). Drive targeted traffic to it via industry newsletters, LinkedIn posts, and personal outreach to known digital publishers. Offer a webinar or whitepaper in exchange for sign-up to gauge serious interest. For those who sign up, follow up within days with a conversation to understand their needs and possibly secure a **letter of intent (LOI)** to pilot.

Success Metric: **Number of qualified publishers** (e.g. with >5M monthly visits) that sign an LOI or express explicit willingness to integrate the MVP. We’d set a threshold like “At least 5 publishers with significant traffic commit to a pilot within 90 days.” Hitting this would validate that the pain is acute enough for action. If we only get, say, 1 or none, that’s a red flag indicating either messaging needs adjustment or interest isn’t there yet.

Experiment 2: Simulated AI Crawler Compliance Test

Goal: Test the technical flow and willingness of an AI agent to pay under real conditions, even if on a small scale.

Design: We’ll use our own simple “AI crawler” (or collaborate with a friendly AI startup) to act as the guinea pig. Take a content source that agrees (perhaps one of the publishers from Experiment 1 or even just a dummy site we control). Implement the end-to-end cycle: the publisher site returns a 402 Payment Required with price in header for our crawler bot; the bot’s code then triggers a micropayment (could be on a testnet for now) and on confirmation, gets the content. We simulate this maybe a few hundred times to mimic a crawling session. This can be done in a controlled environment (no real money, or small amounts).

Success Metric: **Technical success rate and latency** – e.g., “95%+ of crawl attempts successfully complete payment and content retrieval within X milliseconds.” We also measure cost: “Average gas fee per microtransaction = \$Y,” confirming it’s in expected bounds. A soft metric: the reaction of the AI startup (if we involve one) – do they find the integration reasonably straightforward? If we encounter repeated failures, long delays, or the integration proves very cumbersome, we know where to improve. This experiment basically de-risks the core tech in practice.

Experiment 3: Collective Blockade Mini-Pilot

Goal: Understand the dynamics of coordinated enforcement on AI behavior or negotiation.

Design: Identify a small group (even 2–3) of content sites in a particular niche (e.g. 3 tech blogs or 3 local news sites) and help them **simultaneously block a selected AI crawler** for a short period, with the option to “pay via our system” for access. Even if our payment system isn’t fully live, we can simulate it by having them serve a notice or require contact for access. The focus is on an AI that actually fetches content regularly – for instance, pick an open-source model’s crawler or a known startup’s bot. Over a 2-week period, have those publishers deny that bot and see what happens: Does the AI operator reach out for access? Does their bot attempt more aggressive evasion? Or do they simply pause crawling that content?

We will monitor via server logs and perhaps directly ping the AI company for comment if possible.

Success Metric: AI response outcome. The ideal “success” (counter-intuitively) would be that the AI vendor notices and reaches out or agrees to work with us to regain access – proving that blocking can force engagement. Even a response like “we noticed and we’ll comply with your terms” would be huge validation. If instead the bot just keeps trying or shifts IPs (evasion), that tells us enforcement needs to be stronger and AI companies won’t come voluntarily – a valuable insight for strategy (we’d double down on needing large coalitions). This experiment is lean if we find willing publishers because it’s essentially configuration and observation.

Prioritization: 1. **Experiment 1 (Publisher LOIs)** – highest priority, as it tests market demand directly. Without publisher buy-in, nothing else matters. It’s also cheap (mostly marketing hustle) and yields potential pilot partners. 2. **Experiment 2 (Tech compliance test)** – second priority, to ensure our core engine works. This could run in parallel with 1, since devs can do this while bizdev does outreach. It validates feasibility and uncovers any glaring tech bottlenecks early. 3. **Experiment 3 (Collective blockade)** – third priority, as it’s a bit more involved and depends on having a few willing publishers (which Experiment 1 might provide). It gives deeper strategic insight into AI behavior, but if 1 and 2 both underperform, we might pivot or hold off on 3. If 1 and 2 go well, 3 becomes crucial to plan the go-to-market approach (do we need a big bang blockade to force adoption, or will one-by-one integration work?).

Each experiment is low-cost: #1 maybe costs a few hundred dollars of marketing spend and time, #2 just developer time and maybe negligible testnet fees, #3 mostly coordination effort. Yet each yields high informational value about demand, technical viability, and strategy for catalyzing the network.

11. Verdict & Recommendation

After rigorous analysis, the **verdict is to put this venture on the Watch-List (with intent to invest pending milestone proof)** rather than immediately invest or pass outright.

Recommendation Memo:

The Pay-Per-Crawl protocol tackles a compelling problem – the misalignment between content creators and AI companies – with a bold solution that could reshape the digital content economy. Our due diligence finds that **the pain point is real and growing**, especially for publishers, and the timing is favorable with regulatory winds at the back ^{28 81}. The proposed solution demonstrates technical feasibility (micropayments via blockchain at negligible cost ⁵¹) and some early validation (major industry players like Cloudflare and TollBit are already pursuing similar concepts, which is a form of market endorsement).

However, we also identified several **unproven assumptions and execution risks** that prevent a full green-light “Invest” decision today: - **Network Coordination Risk:** The model’s success hinges on broad adoption – it’s a two-sided marketplace facing a chicken-and-egg start. There is a **cold start problem**: if not enough premium publishers enforce payment, AI companies won’t join, and vice versa ⁸². This requires extraordinary go-to-market execution and possibly external catalysts (legal or collective action) to overcome. - **Behavior of AI Giants:** Key AI firms still benefit from the status quo and have shown reluctance to pay unless pressured. Our research shows they’ll likely only participate when “forced” by publisher blockades or legal mandate ^{69 83}. Betting on a behavior change without clear leverage is risky. - **Competitive Response:** Incumbents are moving. Cloudflare’s solution, while centralized, is already in beta with big publishers ¹⁶. If they gain traction quickly by leveraging their network, they could lock in the market before an open protocol reaches critical mass. The startup’s decentralization edge is compelling but

will only matter if it can actually scale an ecosystem before incumbents close the window ⁶⁰. - **Legal Uncertainty:** The legal framework (treating payment + tokenized license as contract) is innovative and seems compliant with e-signature laws ⁵⁵. Yet it remains **untested in court** – a judge has never seen a “Crawl NFT” contract. This is a latent risk; a negative legal outcome or just fear of one could slow adoption. That said, no red flags emerged that it’s invalid on its face – just prudent caution.

Given these points, our recommendation is to **closely watch and engage** with the company through its next phase rather than committing full investment immediately. We propose setting **clear milestones** that, if met, would trigger an investment: - Demonstrate a **successful pilot** with at least one recognizable publisher-AI transaction (even small scale) showing the concept working end-to-end. - Secure some form of **commitment from a cohort of publishers** (for instance, letters of intent from a group that represents, say, 100M+ monthly pageviews in aggregate). This would signal that the cold start might be solvable. - Show tangible progress on the **regulatory acceptance** front – e.g. being referenced in an industry standard or being chosen for a sandbox program under the AI Act compliance efforts.

If the startup meets these milestones in the next 6–9 months, we would be inclined to invest, as that would significantly de-risk the proposition and likely come right as the market inflects (better to invest on confirmed momentum than on speculation in this case).

In the meantime, we will add the startup to our **Watch-List** and maintain close communication. We can offer assistance (introductions to potential pilot customers, advice on experiments as outlined) to help them hit those proof points. Essentially, we want a front-row seat to gauge if they convert early promise into actual traction.

Forwardable Memo for Partners:

Pay-Per-Crawl Protocol – Recommendation: WATCH-LIST (High Potential, Needs Proof)

We’ve conducted deep diligence on the Pay-Per-Crawl concept as presented by the founders. We find the problem – uncompensated AI scraping – to be very real, costing publishers billions and driving urgent behavior (73% blocking bots ¹³). The solution proposes a paradigm shift: make AI companies **pay per access** via an open, decentralized standard. Technically, it’s ambitious but feasible with modern crypto tech (stablecoin micropayments <\$0.0001 each ⁵¹). Strategically, it could establish a new value exchange on the web that aligns with impending regulations (the EU AI Act will push AI toward licensed data ²⁸).

However, at this juncture the model is **unproven in the market**. It faces a classic marketplace bootstrapping challenge and entrenched interests used to “free” data. Early competitive moves by Cloudflare and others confirm the market but also raise the bar for success ³. The founding team has strong vision and domain insight, but will need to demonstrate initial adoption and willingness-to-pay on both sides to justify an investment.

We recommend monitoring the company closely rather than immediate investment. Specifically, we look for evidence in the next 6–9 months of: (a) a cohort of publishers onboard (or ready to onboard), and (b) successful paid crawls executed, even in pilot form. Should those occur, this venture could rapidly scale and become the standard for AI data licensing – at which point an investment would be highly attractive. If those signals don’t materialize, the risk remains that the concept falls short of critical mass or is overtaken by a centralized alternative.

In summary, Pay-Per-Crawl is a **high-upside, high-risk** play. It addresses a pressing pain with a clever solution and could define a new market standard (best-case, a multi-billion-dollar platform opportunity). But it must navigate coordination challenges and prove its model in the real world. We propose to **keep it on our radar**, offer support to hit key milestones, and be ready to move quickly on funding when we see validation. This measured approach maximizes our option to invest in a winner without prematurely committing to an unproven bet.

1 8 13 24 25 26 27 63 Cloudflare's Pay Per Crawl for making money with No Traffic: The Data-Driven Disruption of AI's Free Content Buffet (And What It Means for SERP Dominance) - MonetizeMore
<https://www.monetizemore.com/blog/cloudflare-pay-per-crawl/>

2 4 5 11 12 15 23 29 54 81 Cloudflare Wades into the Battle Over AI Consent and Compensation | TechPolicy.Press
<https://www.techpolicy.press/cloudflare-wades-into-the-battle-over-ai-consent-and-compensation/>

3 16 19 22 32 33 56 57 61 83 Cloudflare launches a marketplace that lets websites charge AI bots for scraping | TechCrunch
<https://techcrunch.com/2025/07/01/cloudflare-launches-a-marketplace-that-lets-websites-charge-ai-bots-for-scraping/>

6 17 18 Atlantic, Vox Media ink licensing, product deals with OpenAI
<https://www.axios.com/2024/05/29/atlantic-vox-media-openai-licensing-deal>

7 9 10 14 20 30 46 49 50 62 64 65 Here are the biggest misconceptions about AI content scraping
<https://digiday.com/media/here-are-the-biggest-misconceptions-about-ai-content-scraping/>

21 38 39 58 66 The Content Scrape Economy: How TollBit is Rewriting the Rules of AI-Driven Publishing Revenue
<https://www.ainvest.com/news/content-scrape-economy-tollbit-rewriting-rules-ai-driven-publishing-revenue-2506/>

28 51 52 55 59 60 69 70 75 76 77 82 Pay-Per-Crawl Protocol Viability Analysis .docx
<file:///file-122ztZ5yZiMjjMLXHeFozQ>

31 53 71 72 73 74 Introducing pay per crawl: Enabling content owners to charge AI crawlers for access
<https://blog.cloudflare.com/introducing-pay-per-crawl/>

34 35 36 37 44 67 68 TollBit raises \$24M series A to ease AI-media data tensions
<https://www.axios.com/2024/10/22/ai-startup-tollbit-media-publishers>

40 41 80 TollBit's New Tool Helps Publishers Save Money From AI Bots Scraping Their Content
<https://www.adweek.com/media/tollbits-publishers-monetize-ai-bots/>

42 43 45 How to Control and Monetize AI Bot Traffic Using Fastly and TollBit | Fastly
<https://www.fastly.com/blog/how-to-control-and-monetize-ai-bot-traffic-using-fastly-and-tollbit>

47 48 Latest AI Breakthroughs and News: May, June, July 2025 | News
<https://www.crescendo.ai/news/latest-ai-news-and-updates>

78 Article 10: Data and Data Governance | EU Artificial Intelligence Act
<https://artificialintelligenceact.eu/article/10/>

79 AI Act | Shaping Europe's digital future - European Union
<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>