Anti-Scraping Countermeasures & Circumventions

CDT Private Roundtable: Preserving an Open Internet in the AI Age

Nick Sullivan · February 10, 2026

The Old Deal Is Breaking Down

Crawlers consumed bandwidth. In return, they sent traffic back.
Publishers ate the cost because the return was worth it.

AI crawlers broke that deal.

CRAWLER	PEAK CRAWL-TO-REFERRAL RATIO (2025)	
OpenAI	**3,700 : 1**	3,700 pages fetched per visit sent back
Anthropic	**500,000 : 1**	Half a million pages per referral

Cloudflare, Jan-Jul 2025. iFixit: 1M hits/day from Anthropic. Freelancer.com: 3.5M requests in 4 hours.

OPERATIONAL COST Bandwidth, server strain, infrastructure burden

IP EXTRACTION Content in training data without permission or compensation

A hobbyist forum cares about staying online. A news publisher cares about its journalism training a competitor. An open-knowledge project may welcome reuse but still want attribution. Not every site cares about both.

Defenses vs. Circumventions

DEFENSE	CIRCUMVENTION	COST
robots.txt	Ignore it (only 30.7% of bots comply with disallow-all)	Free
IP blocking	Residential proxy rotation (190M IPs, $10-15/GB)	Low
Rate limiting	Distributed server-swarms	Low
CAPTCHAs	Solving services ($1-3 per 1,000)	Low
User-agent checks	Browser impersonation	Low
JS challenges	Headless browsers (Google's SearchGuard: "millions of dollars")	Med
Fingerprinting	Anti-detect browser frameworks	Med
Behavioral analysis	Human-like timing, mouse simulation	High
All of the above	*Scrape from Google's cache instead of the site itself*	*Med*

Rows 1-8 address **operational cost.** Row 9 is the **IP problem**: content extracted without touching your servers.

Not All Scrapers Are the Same

1. Compliant

Registers user-agent, follows robots.txt, respects rate limits and HTTP error codes

High velocity, lowest cost

IP extraction

Googlebot, GPTBot

2. Aggressive

No user-agent, ignores robots.txt, brute-forces at scale

Can crash small sites (unintentional DoS)

Operational cost

ByteDance Bytespider
25x faster than GPTBot. 1.4M hits/day.

3. Evasive

Residential proxies, spoofed fingerprints, human-like timing

Low velocity, high cost ($10-15/GB)

IP extraction

Perplexity
Scraped Google's cache when blocked. Reddit caught them with a honeypot.

Most defenses only catch **category 2**. Categories 1 and 3 require different approaches.

Tracing Scraped Content

Google v. SerpApi (DMCA §1201, 2024)

Reddit v. Perplexity (DMCA + trespass, 2025)

LinkedIn v. Proxycurl (CFAA, 2025)

Enforcement is possible. But it needs **proof**.

Operational **Measurable today**

- Bot traffic volume (WAF/CDN logs)
- Known crawler identification
- Rate limit violations

IP extraction **Not measurable**

- Content in a training dataset?
- Which session extracted what?
- Which models used my content?

Honeypots (proven: Reddit)

Watermarking (per-session fingerprints)

Canary tokens (synthetic data)

Membership inference (statistical tests)

Early-stage. Need testing, refinement, and court validation.

Three Paths Forward

Standards & Norms

Preference vocabulary (IETF AIPREF)

Cryptographic bot authentication

Express restriction *or openness*

Ops IP

Voluntary. Most crawlers already ignore robots.txt.

Legal Enforcement

DMCA §1201, CFAA, contract

Case law establishing precedent

IP

Expensive, slow. Needs proof that's hard to get.

Technical Measurement

Watermarking, canaries, fingerprinting

Detect and attribute extraction

IP

Nascent. Active research, not a product yet.

Standards for the willing.
Law for the identifiable.
Technology for the rest.

Anticipated Questions

How do you allowlist legitimate crawlers (Internet Archive, accessibility tools) without also allowlisting scrapers impersonating them?

Cryptographic bot authentication (Web Bot Auth) lets crawlers prove identity with verifiable credentials, not user-agent strings. Decentralized: no single gatekeeper.

Watermarking, canary tokens, fingerprinting: these are surveillance techniques. How is this different from the tracking infrastructure privacy advocates have fought for a decade?

Watermarks travel with content, not users. No cookie, no device fingerprint, no cross-site tracking. Closer to a photographer's EXIF metadata than adtech surveillance.

Millions of creators use CC licenses to enable reuse. Shouldn't the default be open, with opt-in restriction?

Agree: default should be open. Measurement is opt-in. CC-BY-SA has conditions (attribution, share-alike). Measurement helps CC creators verify those conditions are met.

What do publishers actually do on Monday morning? What's the actionable advice for a small publisher?

Four free steps: update robots.txt (establishes legal standing), check CDN bot analytics, review server logs for GPTBot/ClaudeBot/Bytespider, participate in AIPREF standards work.

Wikipedia wants to be crawled. Your framework assumes publishers want to restrict access. What about the open knowledge commons?

The framework is choose-by-default, not restrict-by-default. Measurement helps open projects too: verify crawlers respect CC-BY-SA attribution instead of stripping it.

AI systems that train on web content also drive traffic back through citations and links. Isn't the crawl-to-referral framing ignoring new forms of attribution?

The measured ratios (3,700:1, 500,000:1) are the imbalance. Search clickthrough rates are collapsing as AI answers replace links. The net direction is clear.

Text watermarking is fragile: paraphrasing, summarization, and tokenization destroy most watermarks. Aren't you overselling tools that don't work yet?

Canary tokens aren't watermarks. A canary is a fabricated fact, not a stylistic signal. Paraphrasing doesn't destroy it. Reddit proved the concept. Zero investment is the wrong answer.

Over 20% of the web is behind Cloudflare. If they decide who's a bot, that's a chokepoint. Your measurement tools add another layer. Who operates it?

Open standards with multiple implementers: the Privacy Pass model. One company proposed it, IETF standardized it, multiple vendors operate it. Measurement should be open-source and publisher-deployed.