CRAWL4AI

Open-Source LLM-Friendly Web Crawler & Scraper

GITHUB TRENDING

🏆 **#1 Repository Of The Day**

⭐	🔀	💬	📦
55K	**5.4K**	**4K+**	**900K+**
GitHub Stars	Forks	Discord Members	Monthly Downloads

Download PDF

💡 The Thesis: When Capability Becomes Abundant

> *"When a capability becomes abundant, its complement becomes valuable"*

⚡ Electricity ↓ **Appliances**	🌐 Internet ↓ **Content**	☁️ Cloud ↓ **SaaS**	🤖 AI ↓ **Data**

Bottleneck shifted: Models → Data | Code → Context

Whoever controls the flow of context, controls the flow of intelligence.
The real platform opportunity isn't AI models—it's the data infrastructure layer beneath them.
Crawling is becoming inevitable infrastructure.

→ Control the flow. Own the platform.

💥 New Catalysts Driving Demand

AI Boom RAG, agents, fine-tuning need domain data	**Dynamic Web** JS & CAPTCHAs break old scrapers	**Cloud Costs** API vendors make scale uneconomical
Enterprise Compliance On-prem/hybrid needed for privacy & regulation	**Agent Growth** Real-time, browser-native interactions	**AI-Ready Formats** LLMs need structured, semantic outputs

🛠️ The Origin: Born from Rebellion

📖 My Journey

16	25+	18mo	🔍
First Crawl	**Career**	**Needed Data**	**Searched**
IE4 • Helping dad	Living in this space	LLM-friendly markdown	Looking for solutions

⚠️ The Problems I Found

Outdated Libraries

- ✗ Broke on modern web
- ✗ Couldn't handle JS-heavy sites
- ✗ Not AI-friendly—raw HTML only

FireCrawl

- ✗ Just an API gate—no real code
- ✗ Using LLM for markdown
- ✗ Low quality, high prices

🔥 Rebellion → Action

😤

I hate seeing OSS abused

⚡

Anger → Action
Rebellion mode activated

⚙️

Built Crawl4AI
Weekend • No LLM • Self-hosted

What started as a library became a missing layer in the AI stack

🚀 The Vision: Data Ownership for Everyone

Data + Context + Infrastructure = Platform

👨‍💻 Developers

Clean, private, affordable data pipelines

🚀 Startups

Own your stack • Build your moat

🏢 Enterprises

Control private data • Can't send to APIs • Need infrastructure

🌍 Everyone

Personal AI future • Data accessibility for all

Steve Jobs

"Everyone should learn to code"

I say:

"Everyone should own their data"

Crawl4AI isn't just a tool—it's the infrastructure.
Control the flow of data. Own the platform. Build the future.

❤️👴 This is Why We Build

Yuran is 70 years old. He came to our Discord to learn Crawl4AI. His mission? Crawl all recipes for cooking food with sausage.

Why? He wanted to build an AI app to teach his grandchild how to cook.

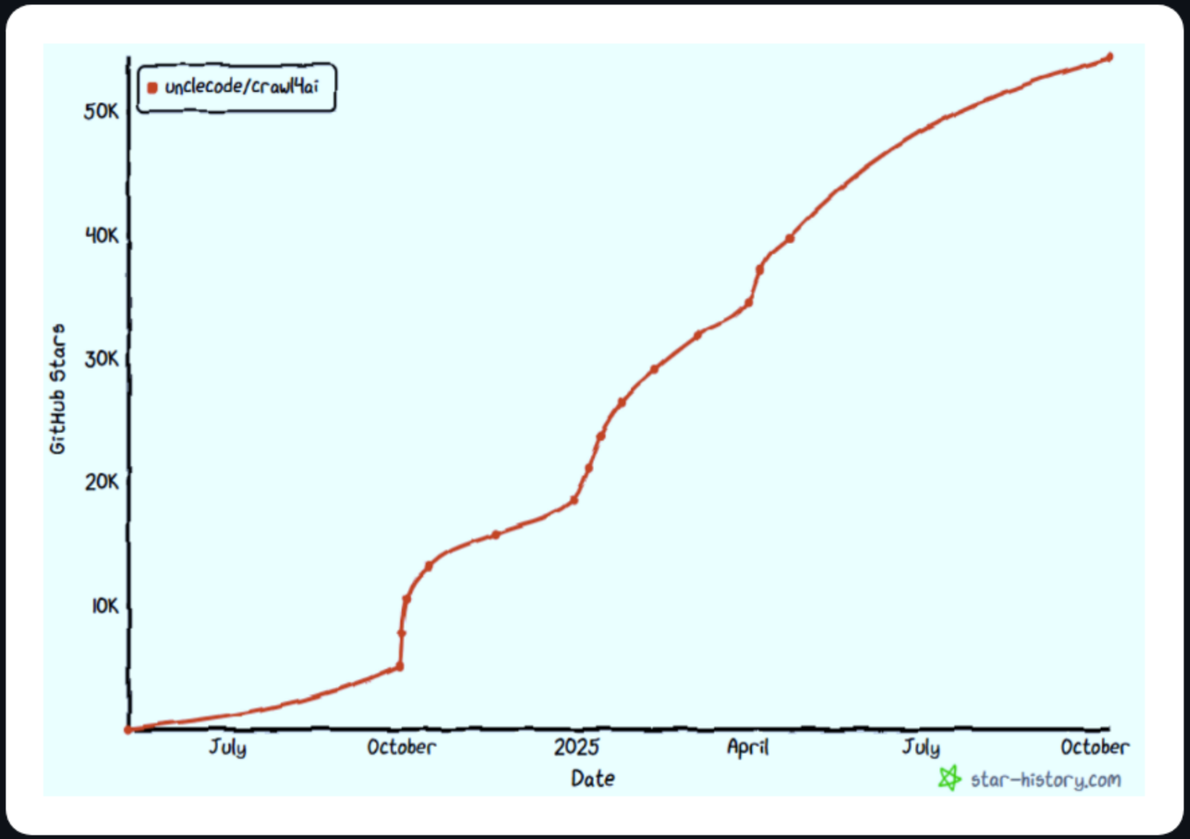
I asked Nasrin, one of our core developers, to guide him through everything. That's what we always do—*from 10-year-olds to 70-year-olds, coders and non-coders.*

His message: "You can't imagine how grateful I am for your help. So many forums have no patience with newbies. But you have put so much time on me. 1000 stars in my book."

This is the Power of True Accessibility

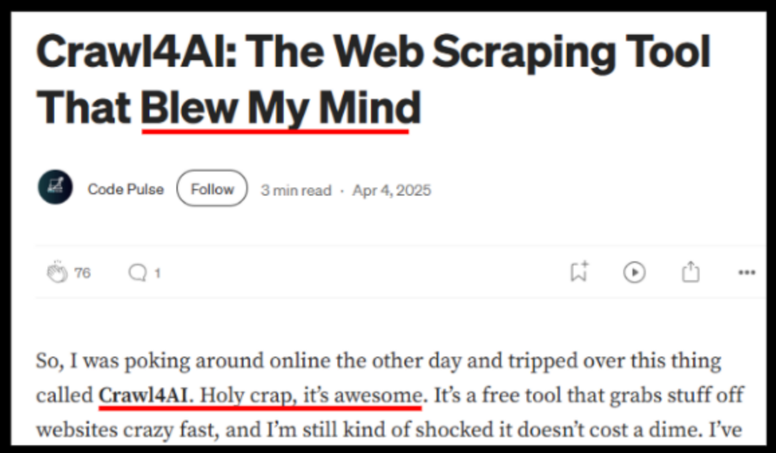
When I say "everyone should own their data," I mean **everyone**. Not just developers. Not just enterprises. Everyone.

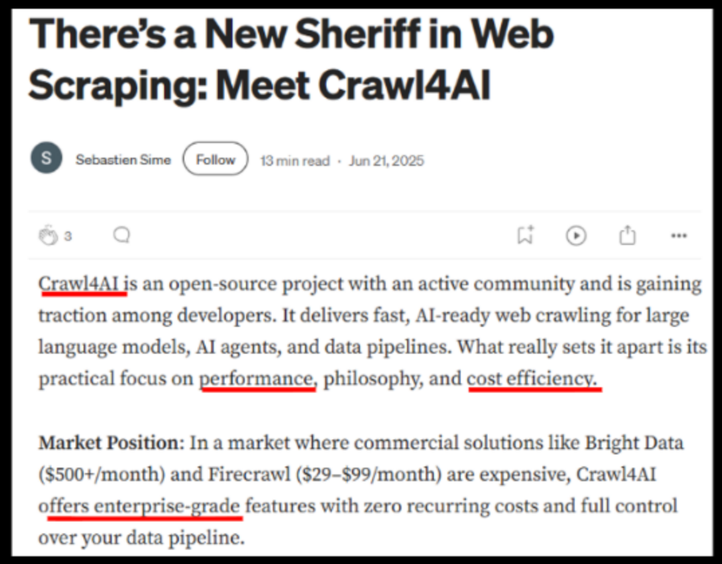
🚀 🤖 Who we are

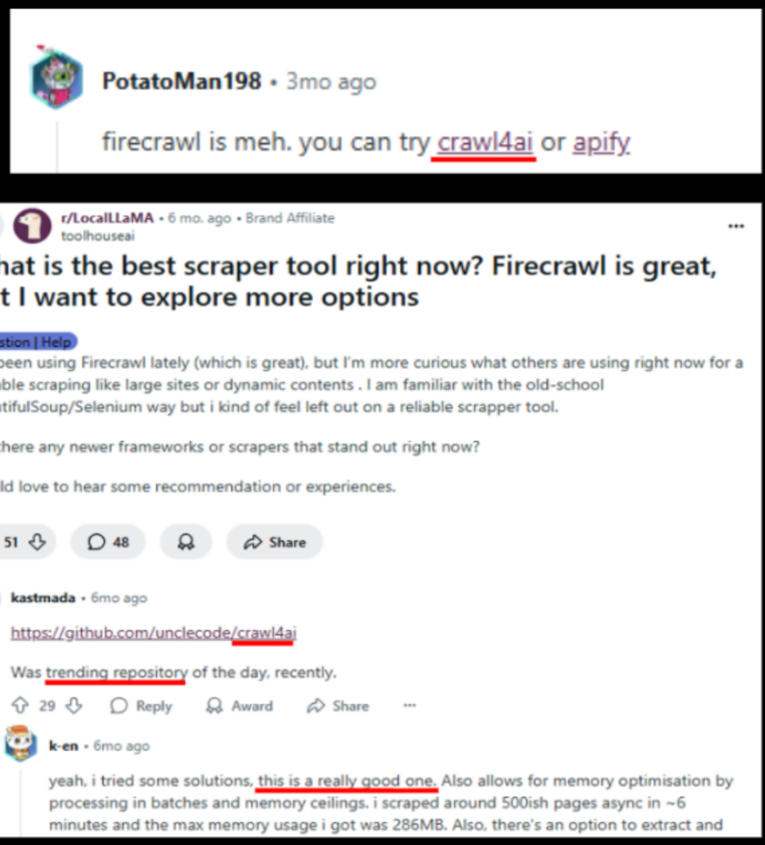


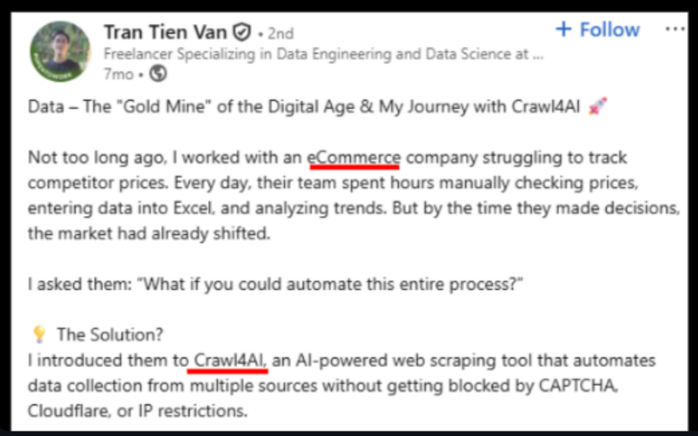
- 📌 **#1 GitHub crawling repo** — 55K ⭐, 900K+ monthly downloads, 2.6K dependents
- 💊 **Born from frustration** with Firecrawl's "open source" limits
- ⚡ **Blazing-fast**, AI-ready crawling for LLMs, agents & data pipelines
- 🔧 **Open source, flexible, and built for structured extraction**
- 🤝 **Thriving dev community** with same-day releases & active Discord

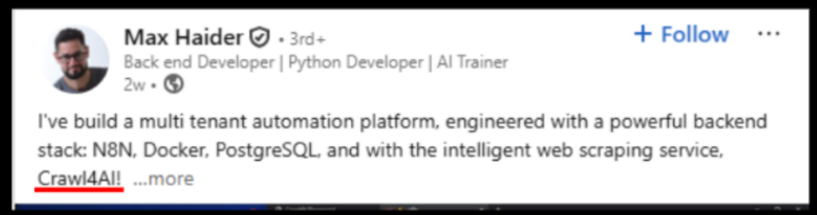
🚀 🤖 Beloved by the Community

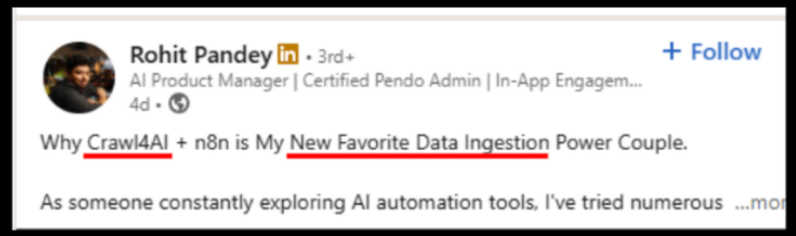




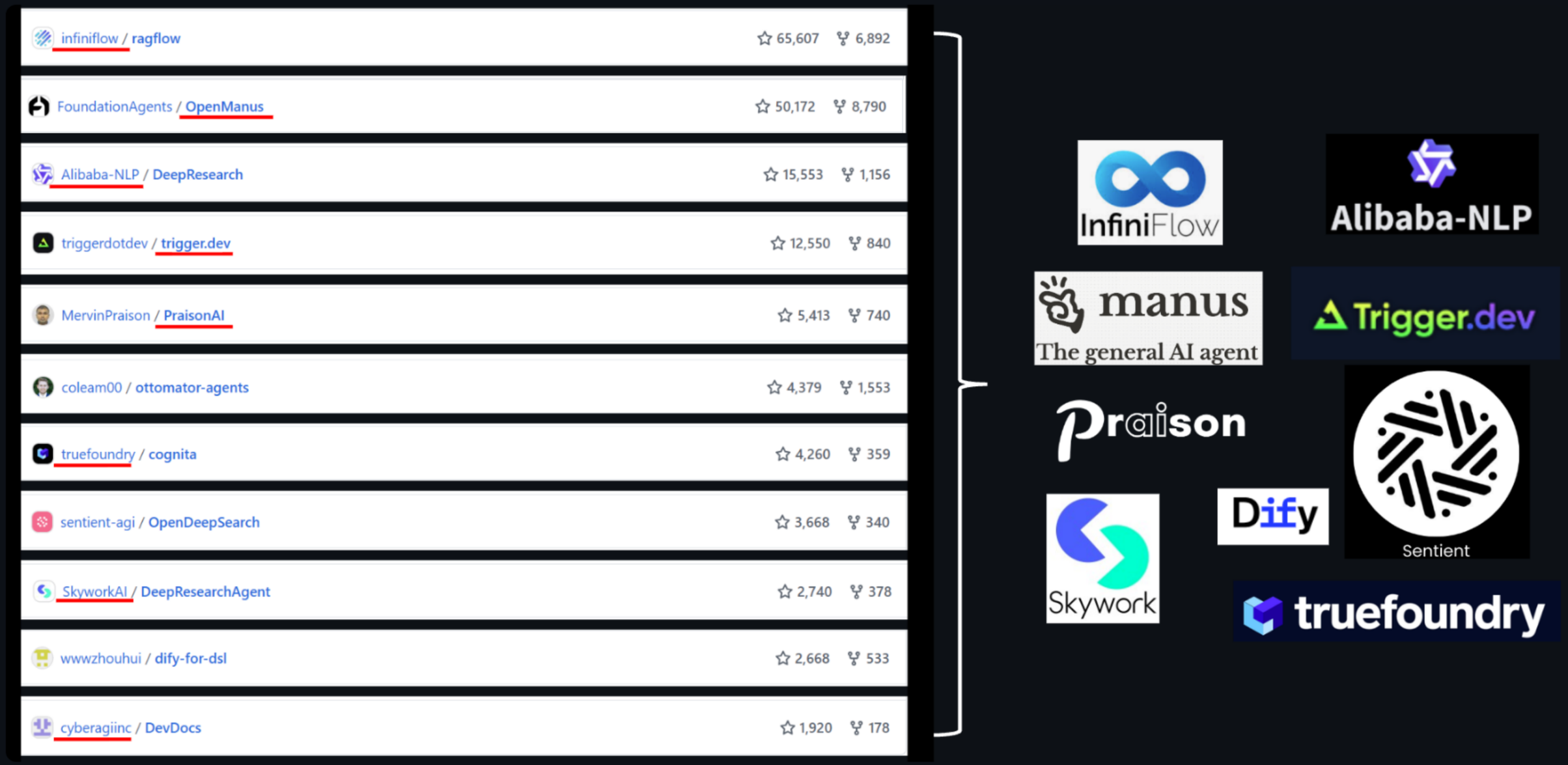








🚀 🤖 Teams using Crawl4AI



🚀🏗️ Infrastructure, Not Tools: Why Crawl4AI Wins

Our Perspective on the Market

- We're not building **"another crawling tool"** — we're building the **infrastructure layer** for AI data
- Firecrawl, Apify, Scrapy? Those are **tools**. We're becoming the **platform**
- Think: **AWS for web crawling**, not "competitor to tool X"

Our community even makes comparisons for us—we don't need to.

📊 See Community Comparison

🔧 Self-hosted & Enterprise-ready

Others lock you into their cloud. We give enterprises control—critical for compliance, security, data sovereignty.

💼 **Business Value: Enterprise contracts + Premium pricing**

⚡ Much Faster / Superior Performance

Speed = cost savings at scale. When crawling millions of pages, faster execution means dramatically lower infrastructure costs.

💰 **Business Value: Operational advantage + Lower churn**

🧠 AI-native Outputs & Multimodal Roadmap

We generate structured, LLM-ready data WITHOUT requiring LLMs (unless semantic needs demand it). Others depend heavily on LLMs = massive cost overhead.

💎 **Business Value: Huge cost savings + Future-proof features**

🏗️ Infrastructure Vision: Commodity → Decentralized

We're not just software—we're building the browser infrastructure layer. Decentralized browser network already in testing.

🚀 **Business Value: Platform economics + Defensible moat**

🤝 Community-driven, Rapid OSS Iteration

55K developers = fastest feedback loop in the industry. We ship what the market actually needs, not what we think they need.

⚡ **Business Value: Product-market fit velocity**

🎯 Because Shaun & Unclecode Are a Hell of a Team

Shaun: Sharp business mind. Unclecode: Relentless builder. Together? Hell of a team. This partnership is why I'm here.

🔥 **Business Value: Perfect founder-investor fit + Accelerated growth**

We're the defacto standard for AI web crawling.
OSS traction proves product-market fit. Technical advantages make it a billion-dollar infrastructure business.

> **Note:** While I'm not focused on feature comparisons, a detailed technical comparison table is available in the appendix for those interested in evaluating Crawl4AI purely from a crawling perspective.

🚀 🤖 Our Market Opportunity

🌐 Market We Serve

Complete web data infrastructure for AI:

- **Core infrastructure:** Crawling, extraction, browser automation, anti-bot, proxies
- **SERP/Data APIs:** Beyond search engines—social media, reviews, pricing, alternative data feeds
- **LLM-ready data:** Structured, semantic outputs AI models can directly consume without processing

🎯 Primary Buyers

- AI-native companies building agents and RAG systems
- Data brokers and infrastructure vendors
- Enterprises across finance, e-commerce, research, healthcare

💰 Market Size

Scraping / Proxy: **$1-3B** Headless / Automation: **$1B+**

AI Data Infrastructure: **$XB+** (multi-billion & accelerating)

📊 **Total: $5-8B today & expanding fast with AI agents & multimodal workloads**

🚀 Four Paths to Scale

1 ✓

Open-Source Adoption

DONE: 55K GitHub stars, massive developer community, proven product-market fit

2

API + Enterprise

Convert adoption to revenue via SaaS model when users need more capacity

3

Infrastructure / BaaS

Platform economics, replace expensive cloud scraping spend

4

Decentralized + Models

Multi-billion moonshot through network effects and high-margin AI services

We've already won developer hearts with OSS—now we monetize when they scale.
Our mission: make sure everyone can crawl, then capture revenue when they need enterprise capacity.

🚀 🏢 Parallel Track 1: Enterprise Monetization

Three Connected Phases: Each builds revenue & capability for the next

1 Monetize What We Built

Turn OSS traction into revenue engine

🔧 CORE FOUNDATION

- **55K stars** → proven product-market fit
- Dual-licensing + auth + metering

⚡ HOSTED API

- **Crawl4AI-as-a-Service** (crawl/extract/map)
- Freemium → Premium (proxies, stealth, concurrency)
- SDKs (Python/JS/Go) + dashboard + logs

💰 **OUTCOME**

Generate **first revenue stream** from API usage → Fund product development & team growth

2 Expand Who Buys & How Much They Pay

Open gates to enterprises & non-technical buyers

🎯 VERTICAL SOLUTIONS

- **Industry pipelines:** e-com price monitoring, lead-gen
- **No-code UI** for non-technical buyers
- Plug-and-play data products

🏢 ENTERPRISE SCALE

- SLAs, security, RBAC, multi-tenancy
- Global proxies + orchestration + monitoring

📈 **OUTCOME**

10x addressable market + capture big contracts → Massive ACV increase & enterprise validation

3 Create Platform & Own Ecosystem

Beyond product—become the crawling infrastructure layer

🔧 INFRASTRUCTURE / BAAS

- **Become AWS/Cloudflare for crawling**
- Managed browser pools as core AI infrastructure
- Global, distributed, scalable data platform

🧠 DATA → MODEL SERVICES

- Fine-tune LLMs on crawled data (high-margin, sticky)
- Synthetic dataset generation
- Private model training on company

🔗 STRATEGIC INTEGRATIONS

- Integrate with Snowflake, Databricks, BI/ML platforms
- **Default data layer** for AI infrastructure
- Distribution moat through platform partnerships

🚀 **OUTCOME**

Ecosystem ownership → Platform economics + strategic positioning = Multi-hundred million valuation

🎯 The Compounding Effect

Phase 1 proves the business model and generates capital → **Phase 2** expands market reach and validates enterprise value → **Phase 3** creates irreplaceable infrastructure and ecosystem lock-in → Result: We don't just sell a product—we own the crawling ecosystem.

🚀🔬 Parallel Track 2: Moonshot / Skunkworks

These are moonshot R&D initiatives—experimental, high-risk, high-reward. Plans will evolve based on real market feedback. Focus areas: decentralized infrastructure, mobile/edge crawling, and multimodal capabilities.

🌐 Decentralized Browser Network

✓ TESTING

"Browser Net" - Orchestration platform for asynchronous crawling infrastructure

- **Distributed nodes** provide browser & crawl capacity
- Token / rev-share incentives + residential IP diversity
- Coordinated orchestration for large-scale asynchronous jobs

💡 **WHY THIS MATTERS**

Unlocks capacity & scalability: If successful, removes infrastructure bottlenecks and enables unlimited scale without proportional costs. Network effects create defensible moat.

📱 Mobile / Edge Crawling

Real-world data from mobile devices

- **Android / iOS nodes** for real-world data
- Mobile devices expand global coverage
- Access app data & mobile-only experiences

💡 **WHY THIS MATTERS**

Mobile-first world: Massive markets only accessible via mobile. Unlocks app data, regional content.

🖼️ Multimodal Crawling

Beyond text—images, audio, video extraction

- **Extract faces, logos, A/V segments** at crawl time
- AI-optimized, vertical-specific datasets
- Computer vision & multimodal model training

💡 **WHY THIS MATTERS**

AI is going multimodal: Position at the data layer for vision, audio, and video models.

🏗️ Commodity Browser Infra

Custom Chromium for scale

- **Custom Chromium** for high-concurrency crawling
- Low-cost, non-AWS infrastructure for scale
- Shared layer for developers & competitors

💡 **WHY THIS MATTERS**

"Picks and shovels": Become the infrastructure layer everyone uses. Platform economics with massive TAM.

🎯 **Adaptive Strategy**

These initiatives are R&D experiments, not commitments. **Decentralized network is already under testing—proving feasibility.** Mobile and multimodal will be refined based on enterprise feedback and market demand. We'll double down on what works and pivot away from what doesn't. Each could be a standalone billion-dollar business.

🤝🚀 Why Partner? *(Not Why Fundraise)*

WHY I'M FUNDRAISING:

Not to prove the concept—**55K stars did that.**
Not to survive—I have runway and clear monetization.

I'm fundraising to accelerate the inevitable:
Someone will own the data infrastructure layer for AI.
It should be us.

📊 TYPICAL AI INVESTMENTS	✓ THIS OPPORTUNITY
API wrappers with no real tech	55K organic stars, real infrastructure
Bought GitHub stars & fake traction	Authentic community, daily adoption
"Open source" that's just marketing	True OSS, self-hostable from day 1
Founders proving concept	Founder who already proved it

💪 What I Bring:

- Technical moat & deep domain expertise
- Community trust & organic adoption
- Execution speed & bias to action
- Infrastructure thinking, not product thinking

I can build a stable business alone.

🎯 What You Bring:

- Thought partner for GTM strategy
- Strategic guidance for platform plays
- Capital for moonshot R&D
- Long-term thinking, not quarterly pressure

Together we build global infrastructure.

🎯 The Choice

For VCs: Back authentic traction vs. manufactured hype.
Control the flow of context and intelligence—or watch someone else do it.

For me: Partner with believers who understand infrastructure plays.
I'm not looking for money—I'm looking for co-pilots.

This is the moment. Let's build infrastructure.

💸 The Ask

$3-5M at $25M post | Long-term partnership with dreamers

CRAWL4AI

Open-Source LLM-Friendly Web Crawler & Scraper

GITHUB TRENDING

🏆 **#1 Repository Of The Day**

⭐	🔀	💬	📦
55K	**5.4K**	**4K+**	**900K+**
GitHub Stars	Forks	Discord Members	Monthly Downloads

🚀 🤖 Busines Roadmap

Fundraise	OSS Phase	Timeline	Business Milestone	Key Roles / Capabilities	Tech Stack Focus	Team Size	OPEX (USD/yr)
Seed ($3 – 5 M) OSS Foundation Open Core shift	🧩 0. OSS Foundation & Self-Hosting Simplification	0 – 4 mths	• Self-hosting (one-click deploy, Docker) • Lay foundation for dual-license model. • Build trust with developer community	+1 Founder-Engineer (Python core), +1 Crawler specialist, +1 DevOps (Docker orchestration).	Python (core library), Playwright / Puppeteer, Docker / Compose / Helm, CI/CD.	3	< $0.4 M
	☕ 1. Hosted API Commercialization (“Starbucks Business”)	4 – 10 Mths	• API layer (Starbucks business). • Focus on reliability + efficiency • Serve developers & SMBs	+ 2 Backend Eng (Node.js / TypeScript), + 1 Frontend (React.js), + 1 DevOps / SRE.	Node.js / TypeScript (API), Redis, Postgres (Supabase), React.js UI, Kubernetes for scaling.	6 – 8	≈ $1.2 M
	🏢 2. Enterprise Licensing / On-Prem Commercialization	10 – 15 mth	• Enterprise commercialization • Target data-intensive enterprise clients • Clear cost superiority vs peers	+ 2 Backend Eng (enterprise infra), + 1 QA /Test, + 1 Support / Sales Eng, + 1 Customer Success.	Python + TypeScript stack; secure deployment (SAML/OAuth), Kubernetes, Terraform, CI pipelines, private package management.	10 – 12	≈ $2 M
Series A ($15 – 20 M) Scale infra & GTM	🧠 3. Infrastructure Play — Browser-as-a-Service (BaaS)	15 – 25 mths	• Infra play: Browser-as-a-Service • backbone for Crawl4AI + 3rd-party vendors • Strengthen margin profile	+ 2 Crawler / Infra Eng (browser orchestration), + 1 Infra Architect, + 1 ML Eng (optimize runtime), + 1 Biz Dev (Infra partnerships).	Rust / Go modules for browser orchestration, Node.js controllers, GPU / CPU autoscaling, load balancing, monitoring (Grafana/Prometheus).	14 – 16	≈ $2.8 M
	🌅 4. Multimodal & Advanced Crawling	25 – 36 mths	• Multimodal / advanced crawling • Mobile & edge crawling (Android/iOS)	+ 1 ML / CV Eng (vision/audio extraction), + 1 Mobile Eng (Android / iOS crawlers), + 1 QA Automation for multimodal.	Python (CV / audio), TensorFlow / OpenCV, Mobile SDK integration, LangChain for structured output.	18 – 20	≈ $3.2 M
Series B ($30 M +) moonshot infra	🌐 5. Decentralized Browser Network + Data / Model Services	36 – 48 mths	• Decentralized browser network. • Data/model services (fine-tuning LLMs) • Dual-track moonshot with enterprise track	+ 2 Blockchain / P2P Eng, + 2 Data / ML Eng (fine-tuning), + 1 Platform PM, + 1 Partnerships Mgr.	Rust / Go for P2P networking, Python for data & model services, LangChain, Hugging Face, tokenization infra.	22 – 25	≈ $3.8 M

🚀 🤖 Meet our neighbours (not competitors)

⚔️ Neighbours	🏷️ Category	💡 Our Distinction / WHY US?
Modern LLM-ready web crawlers	🔥 Firecrawl	🧩 **Truly self-hosted** — ultra-fast, efficient, low-cost, community-driven
Managed browsers, proxy/IP rotation	🧠 Browserless / Playwright / Bright Data	⚙️ **AI-native extraction** — structured / Markdown outputs → future BaaS & decentralized “browser net”
Classic Web-Scrapers	🕸️ Scrapy, Apify, Colly	🤖 **LLM-optimized** — multimodal, one-click self-hosting, on-prem
Data-Enrichment / Lead-Gen Apps	💼 Clay, Apollo, Clearbit	🧰 **Infra, not SaaS** — developer-neutral, enterprise-grade automation library
Proxy / IP Providers	🌐 Bright Data, Oxylabs, Smartproxy	🧠 **Beyond access** — integrated crawler engine + community node roadmap
Site-Specific APIs	🔍 Zyte API, SerpAPI	🕹️ **Full-page automation** — AI-ready, self-hostable, enterprise-scale