The Web Is Forking: What the Agentic Web Means for Local AI Builders

February 21, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Every major infrastructure company is simultaneously building a different piece of a new web — not for humans, but for AI agents. Coinbase built agent wallets (x402 protocol). Stripe built agent payment tokens. Cloudflare auto-converts sites to markdown for agent consumption. Exa.ai built search specifically for agents. OpenAI shipped execution environments. None of them coordinated. The web is forking into human and agent layers on the same infrastructure. If you run local AI, your agents are about to become participants in this new web — and understanding the fork early is the advantage local builders have always had.

:books: **Related:** The 5 Levels of AI Coding · llama.cpp Just Got a New Home · What Open Source Was Supposed to Be · Running AI Offline

Something is happening across the internet right now that doesn't have a single announcement or launch event. Coinbase, Stripe, Cloudflare, Google, OpenAI, Visa, PayPal — they're all building different pieces of the same thing. Independently. Without coordination. Within the same few months.

They're building a second web. Not replacing the one you're reading this on — running alongside it, on the same infrastructure, for a different kind of client. Not humans with browsers. Software that reads, decides, pays, and acts.

The agentic web. And if you build with local AI, you're already part of it whether you realize it or not.

The Fork

Think about what a webpage is designed for. Fonts. Layouts. Hero images. Scroll animations. Cookie banners. A navigation menu with hover states. All of that exists because the client is a human with eyes and a mouse.

Now think about what an AI agent needs from that same page. The text. Maybe a table. Structured data. A price. An API endpoint. Everything else — the CSS, the JavaScript, the 47 tracking pixels — is noise that consumes tokens and context window for zero value.

The web has always served one client type: humans using browsers. What's happening now is a second client type showing up at the same doors, and the infrastructure is being rebuilt to serve both. Not a replacement. A fork.

This isn't theoretical. The pieces are already shipping.

The Infrastructure, Layer by Layer

Money

Coinbase launched Agentic Wallets in February 2026, built on their x402 protocol (named after HTTP 402 "Payment Required"). The protocol has processed over 50 million machine-to-machine transactions. Agents get non-custodial wallets with keys isolated in Trusted Execution Environments — the agent can authorize payments but never sees the private key. Within 24 hours of launch, thousands of AI agents had registered wallets on Ethereum.

Stripe shipped their Agent Commerce Suite with Shared Payment Tokens — scoped, time-constrained credentials that let an agent buy things without ever seeing a card number. Bounded by merchant, amount, and time window. Revocable instantly. Their fraud detection system (Radar) had to be extended with entirely new signal types because decades of fraud heuristics — mouse movement, browsing time, device fingerprints — are meaningless when the buyer is software.

Google announced the Universal Commerce Protocol at NRF 2026 in January, with over 20 retailers and payment companies endorsing it, including Visa and Shopify. PayPal partnered with OpenAI in October 2025 for instant checkout inside ChatGPT. Visa shipped their Trusted Agent Protocol — cryptographically signed HTTP messages that transmit an agent's intent, verified identity, and payment details.

The payment rails for agents are being laid by every major financial infrastructure company simultaneously.

Content

Cloudflare shipped "Markdown for Agents" in February 2026. When an AI agent sends `Accept: text/markdown`, any Cloudflare-enabled site auto-converts its HTML to clean markdown before serving it. The response includes an `x-markdown-tokens` header with the estimated token count, so agents can check whether a page fits their context window before ingesting it.

The token savings are dramatic. One test showed HTML consuming 16,180 tokens while the markdown conversion used 3,150 — an 80% reduction. Cloudflare serves roughly 20% of the web. They just made a fifth of the internet agent-readable by default.

They also support llms.txt — machine-readable site indexes that tell agents what content exists and where to find it. InsiderLLM has had llms.txt deployed for weeks. ChatGPT-User crawls us over 1,100 times per day and recommends our articles in responses. We're already serving the agent web.

Search

Exa.ai built a search engine from the ground up for agents — their own index, their own neural retrieval, their own embeddings. No SERPs, no ads, no "People also ask" boxes. Raw URLs and structured data. Their Websets Pro endpoint scored 94.9% on SimpleQA in their own evaluation.

Speed matters here in ways it doesn't for human search. Brave's agent search returns results in 669ms (per third-party benchmarks). Some competitors take 13+ seconds. When an agent chains three search calls in a single workflow, that latency difference compounds from seconds into minutes. For agentic workloads, search speed is infrastructure, not a nice-to-have.

Execution

OpenAI shipped three primitives that matter: **Skills** (versioned instruction packages — think Docker images for agent procedures), **Shell** (a real Debian 12 Linux terminal agents can execute code in), and **Compaction** (automatic context window management for long-running workflows). Triple Whale reported their agent navigated a 5-million-token session with 150 tool calls without accuracy degradation.

Glean found that a single well-structured Salesforce skill — one set of instructions walking the model through using multiple tools to analyze data — increased accuracy from 73% to 85% and reduced time to first token by 18%. Structure matters more than model size for agent tasks.

You've Seen This Before

In 2007, the web worked on phones. Technically. You could load CNN.com on your iPhone and squint at text designed for a 1024-pixel monitor. It was functional in the way that reading a newspaper through a keyhole is functional.

What followed was a decade-long rebuild. Responsive design. Touch-first interfaces. App stores. GPS-native features. The companies that recognized the fork early — that the new client wasn't just a smaller screen but a fundamentally different interaction model — built the dominant platforms of the next era. Uber, Instagram, WhatsApp. None of them would have existed on the desktop web.

The agentic web is the same inflection. The new client isn't a smaller screen. It's software that reads, decides, pays, and acts. The infrastructure being built right now — agent payment tokens, markdown conversion, agent-native search, execution environments — is the responsive design moment for AI.

Why Local AI Builders Should Care

If you're reading InsiderLLM, you probably run models on your own hardware. You might think the agentic web is a cloud/enterprise story. It's not.

Your agents need these primitives too

If you're building with OpenClaw, mycoSwarm, or any local agent framework, your agents need the same capabilities: reading web content, searching, executing code, eventually transacting. The difference is you control the agent instead of renting one from OpenAI. But the web your agent talks to is the same web. Cloudflare's markdown conversion works for your local agent's HTTP requests the same way it works for ChatGPT's.

llms.txt is the on-ramp

Every site you run — personal projects, documentation, businesses — should have an llms.txt file. It's how agents discover your content instead of relying on Google. We deployed ours and immediately saw ChatGPT referencing our articles in responses. The format is simple: a markdown file at `/llms.txt` that lists your content with descriptions. Ten minutes of work. Disproportionate return.

Security is non-negotiable

Every serious implementation of the agentic web treats the agent as a potential adversary — not a trusted employee. Coinbase isolates wallet keys in hardware the agent can't access. OpenAI sandboxes shell execution in ephemeral containers. IronClaw (the Rust reimplementation of OpenClaw) sandboxes every tool call in isolated WebAssembly.

If you're running local agents with tool access — file system access, API keys, database connections — you need the same mindset. Your agent is software that executes based on probabilistic text generation. It will occasionally do something wrong. The question is whether "wrong" means a bad API call or a wiped database.

The OpenClaw one-click RCE vulnerability wasn't theoretical. Agents with unsandboxed tool access are a security incident waiting for a trigger.

The trust gap is real

Salesforce found 70% of consumers would use AI agents for routine tasks like loyalty point optimization. But adoption drops sharply for higher-stakes decisions — returns, purchases, financial actions. There's a gap between what the infrastructure enables (fully autonomous agents) and what people actually want (agents that check in before doing anything important).

Every security incident — databases wiped by unsupervised agents, the OpenClaw RCE, iMessage disasters from agents that sent messages without confirmation — pushes the trust timeline back. The infrastructure is building toward full autonomy. Human comfort is at "show me what you're about to do and let me approve it."

That gap is the central tension of the next few years. It's also where local AI has an advantage: you set the guardrails, not a platform.

Agents as Economic Entities

Here's where it gets genuinely novel. Agents with wallets that earn, spend, and accumulate capital independently of their creators. That's a category of software that has never existed.

On Polymarket, AI agents are already trading prediction markets — Polymarket themselves confirmed it, and agents now represent over 30% of total trading volume. Some of these agents are explicitly trading to subsidize their own compute costs. The loop is closing: an agent that earns money to pay for the GPU time it needs to keep earning money.

On a long enough timeline, this creates agents that are economically self-sustaining. We're not there yet. But the primitives — wallets, payment protocols, execution environments, search APIs — are all shipping now. The distance between "an agent that can buy something" and "an agent that runs a small business" is shorter than most people think.

The Scam Warning

TikTok is flooded with “turn $50 into $3,000 with AI agents” content. The reality: profitable agent trading requires latency arbitrage on colocated infrastructure, not OpenClaw bots running on your laptop. The margins are razor-thin and the competition is institutional. One developer publicly reported racking up $200 in API fees in two days trying to build an autonomous trading agent.

The agentic web is real infrastructure being built by serious companies. The “passive income from AI agents” pitch is the 2026 version of “mine Bitcoin on your gaming PC.” The infrastructure exists. The easy money does not.

What to Do Right Now

Deploy llms.txt. On every site you control. It’s a markdown file at your root that indexes your content for agents. InsiderLLM’s is here as a template. Ten minutes. Do it today.

Sandbox your agents. If you’re running local agents with tool access, isolate them. Docker containers, restricted file system access, approval gates before destructive actions. Treat your own agent as untrusted code, because it is.

Watch the standards. x402 for payments. Cloudflare’s markdown conversion protocol. Google’s Universal Commerce Protocol. OpenAI’s Skills format (which Anthropic has also adopted). These are forming now. They’ll matter.

Think of your local agents as participants. Your Ollama instance running a 32B model on a 24GB GPU can make the same HTTP requests, consume the same markdown-converted pages, and use the same search APIs as any cloud agent. The agent web doesn’t check your cloud provider credentials. It checks your HTTP headers.

Bottom Line

The agentic web is real. It’s being built simultaneously by Coinbase, Stripe, Cloudflare, Google, OpenAI, Visa, and PayPal — independently, on the same timeline, reaching the same conclusions. The web is forking into human and agent layers running on the same infrastructure.

The hype-to-reality ratio is dangerous. Timelines will be longer than promised. The scams will be loud. Security incidents will set adoption back. Agents won't be buying your groceries next month.

But the infrastructure is shipping now. And local AI builders who understand the fork early have the same advantage they've always had: you own your agents, your data, and your infrastructure. When the agent web matures, you won't be renting access to it. You'll be running it.

That's the whole point of local AI. It always has been.

Source: https://insiderllm.com/guides/agentic-web-local-ai-builders/

Free guides for running AI locally