PitchBook
a Morningstar company

Institutional Research Group

Rudy Yang
Senior Research Analyst,
Enterprise Fintech and
Retail Fintech
rudy.yang@pitchbook.com

pbinstitutionalresearch@pitchbook.com

Published on July 21, 2026

Contents

INDUSTRY RESEARCH

Machine Economy Rising: The Market of Limitless Cognition

Part I of II

Report snapshot

What the machine economy is

- The machine economy is a market in which AI agents discover, transact, and generate value with little to no human involvement.
- Value can be measured from three broad layers: existing software and labor spending rotating to agents, the infrastructure agents consume to operate, and net new value from an agent-only market that did not previously exist.
- The market expands as intelligence becomes cheaper and more capable, allowing agents to do previously uneconomical work and transact with other agents. Early estimates from Forsy put global agent GDP at $36 billion annualized as of July 2026.[1]

Why the machine economy is emerging

- Existing software and labor spending is already rotating toward AI. We estimate about $20 trillion of AI-exposed work, though only about 1% flowing through agents.
- Token volumes are compounding far beyond what human usage can explain. Google processed 3.2 quadrillion tokens in May 2026, up from 9.7 trillion two years prior.[2] OpenAI and Alphabet each process more than 15 billion tokens per minute via direct API access.[3,4]
- Web traffic is shifting toward machines, with bots now accounting for 57.5% of HTML requests. Observed agentic AI traffic is growing quickly, and browser-based agents make up about 71% of observed agentic traffic.[5]
- Machine-facing infrastructure is spreading fast. We queried 12,594 model context protocol (MCP) servers that were indexed as of June 18, 2026, while Stripe reports that 70% of command-line interface (CLI) API requests now come from agents.[6]

- APIs are becoming machine-scale distribution channels. Salesforce processed almost 1 trillion API calls in Q1 2026 alone.[7] Many platforms are also moving to metered pricing, while roughly 25% of developers now design APIs with agents as the primary end consumer.[8]

What stands in the way

- Agents are not yet priced as economic actors because autonomy remains constrained across four dimensions: scope, controls, supervision, and longevity.
- Our scoping exercise of 40 deployed agent products shows that today's agents are still not true autonomous economic actors, with average scores of 3.2 for autonomy and 3.5 for scope on a scale of 1 to 5.
- In live workflows, autonomy typically breaks at three structural points: when information is not machine-readable, when an agent lacks credentials to act, and when a payment requires human intervention.

Why this shift matters

- The gap between today's agents and fully autonomous ones is where new opportunities are emerging. Companies that solve for discovery, access, usage, and payment at agent scale will be best positioned as the market grows.
- Success will become increasingly dependent on service directories, standardized data formats, command-line access, and pricing models that work for agent-scale usage.
- New revenue models will emerge around per-call, per-outcome, and per-execution pricing, especially as agents start buying services from other agents.
- Solving the agentic payments layer is crucial but difficult. Agents need to hold funds, initiate transactions, and operate within defined spending limits without a human having to re-enter the flow.
- There are big open questions around trust, identity, intent, and reversibility. Companies that solve this layer early will have a durable first-mover advantage.

Introduction

This is a two-part report series on the machine economy and agentic payments. Part I examines the emerging machine economy. Part II makes the case for why machine-first payments infrastructure is needed to unlock it.

AI agents are already moving toward greater autonomy. They are increasingly consuming paid services, spawning more subagents, and completing more tasks with less human involvement. Over time, they have the potential to create net new economic value and support a market that can scale at unprecedented speed.

This shift toward an agent-first market may look inevitable, but it will take enormous investment. Key infrastructure gaps still stand in the way. In Part I, we examine what the machine economy may look like, how much value it can create, why it appears inevitable, and, importantly, what is preventing it from arriving sooner. Within these gaps are new emerging opportunities and important signals that companies should be building around now.

In this section, we define the machine economy, outline a framework for measuring its key components, and consider what a conceptually new agent market could look like as it takes shape.

Sketching a market of unbounded cognition

Underneath the rise of agentic AI is a market worth trillions of dollars, forged by autonomous AI agents transacting and creating value at machine speed. Each time an agent calls a model, invokes an orchestration platform, parses the web, taps an API, or transacts on its own, it is participating in a machine-native value chain. In aggregate, this network of interlocking services enables agents to act autonomously and generate economically meaningful output.

This leads to the rise of a broader machine economy, which we define as a new market in which AI agents discover, transact, and generate value with minimal to zero human intervention. In this new economic layer, the relevant market is any digitally addressable workflow where cognition is worth paying for. Agents perform the work inside systems, turning inputs into decisions, actions, and outputs that generate measurable economic value.

However, defining a shift of this scale is difficult because it blurs the old boundaries between software, services, labor, and digital infrastructure. Instead of sizing markets for software and IT budgets alone, the formula evolves to measure how much of the world's cognitive work becomes legible, executable, and billable through agents. Therefore, a key challenge is quantifying a market that, in theory, maps to the demand for intelligence.

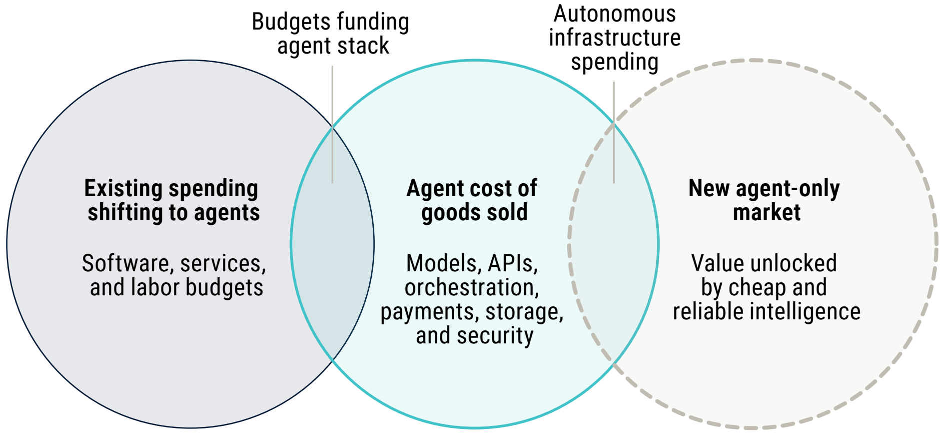
Defining this market is important because it helps specify where spending will go, which layers of the technology stack will matter most, and where new value is likely to accrue. In our view, the machine economy's value stems from three broad layers:

- **Existing spending shifting to agents:** Money that companies already spend on software, outsourcing, and salaried staff, which rotates to agents instead of traditional tools or people.

- **Agent stack revenue:** Spending allocated to the infrastructure that agents run on, which can be thought of as an agent's cost of goods sold. This includes costs on models, APIs, orchestration platforms, databases, storage, and security.

- **New agent-generated value:** Value and economic activity created by agent-only products and services that were previously impossible in human-only workflows. It is made possible when intelligence is cheap, reliable, and able to transact with other agents.

The machine economy



Source: PitchBook

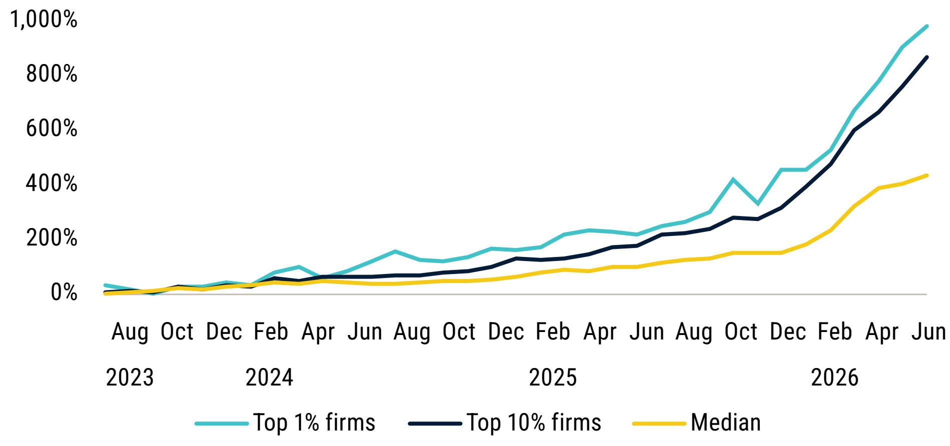
Existing spending shifting to agents

One of the earliest measurable surfaces of the machine economy is the reallocation of existing software, services, and labor spending into agent workflows. Already, money that companies once spent on software seats, outsourcing firms, and salaried staff is now being redirected toward agents that can operate across tools, complete tasks, and deliver outputs within businesses. We have written about this in detail in our previous notes SaaS is Dead, Long Live Sas and Through the Looking Glass: The Race to Build Enterprise AI.

Outcome-based pricing has expanded enterprise software's addressable market beyond the $1.4 trillion annual software spending to encompass a portion of the $50 trillion global knowledge labor market. We estimate the current level of AI-exposed work across these combined markets to be roughly $20 trillion, with only about 1% of that executed by AI agents. This implies that while approximately $200 billion of work today flows through agent-mediated workflows, most software and knowledge-worker spending has yet to be rewired.

Inevitably, this shift will continue as intelligence becomes cheaper and more capable per dollar. Ramp's data from more than 70,000 US businesses shows that the share of firms spending on AI, including large language model (LLM) subscriptions, coding agents, API tokens, and GPU cloud spending, rose to 55% in June 2026 from 42.7% one year earlier.[9] Its data also highlights how rapidly AI spending per employee is rising. The median AI spending per employee increased 435% to $10.70 from a July 2023 baseline of $2.00. For the highest spenders, the increase is even more drastic. From July 2023 to June 2026, the top 1% of firms increased AI spending per employee from $453.68 to $4,883.33, while the top 10% rose from $53.54 to $516.20.[10]

Percentage change in AI spending per employee



Source: Ramp • Geography: US • As of June 2026

Agent stack

The reallocation of spending from software and labor partly funds the agent stack, another highly visible layer of the machine economy. This is effectively an AI agent's cost of goods sold: the runtime software and services an agent actively consumes to operate, including model inference APIs, orchestration and middleware, application hosting, vector and retrieval services, third-party tool APIs, and security and observability. We exclude physical infrastructure beneath this layer—including GPU hardware, datacenters, and dedicated capacity providers—as many of those costs are already embedded in the per-token prices agents pay. Model providers, cloud platforms, and infrastructure vendors are already billing for this layer at scale.

For example, an agent that researches vendors and drafts a recommendation would search through stored contract history, pull live pricing from the web, and call a language model several times to reason across the results and produce a summary. Each step is metered. A simple workflow like this might consume several thousand tokens across model calls, database lookups, and API calls, totaling a few cents per run. Across thousands of agents executing workflows at machine speed, this spend accumulates quickly.

While there is no single quantifier for agent costs, we believe tens of billions of dollars per year are being captured by this rapidly scaling layer. As a 2025 baseline, Menlo Ventures sized AI infrastructure spending at $18 billion, comprised of $12.5 billion for foundation model APIs, $4 billion for model training infrastructure, and $1.5 billion for AI infrastructure covering storage, retrieval, and orchestration.[11]

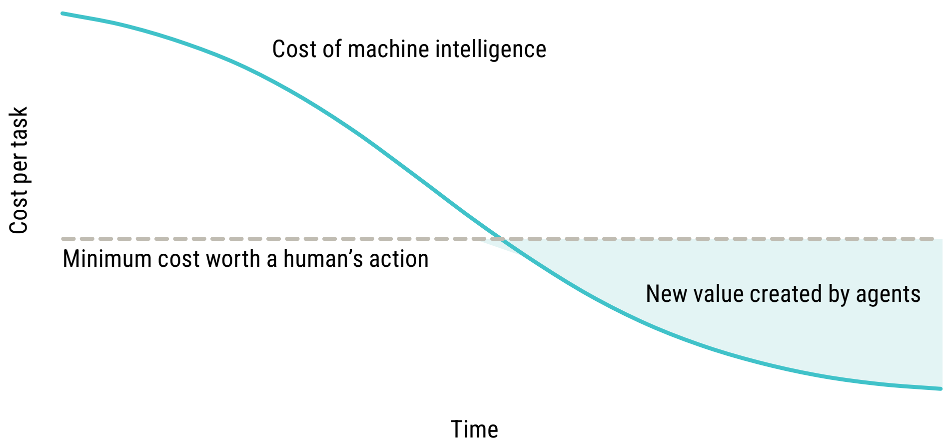
However, run rates for 2026 suggest the stack has already grown several-fold. The two largest model providers, Anthropic and OpenAI, are projected to generate nearly $80 billion in combined annualized revenue alone on a run-rate basis.[12, 13] Not all of this reflects agent activity because both labs also earn subscription and consumer income, but we estimate the majority is increasingly weighted toward enterprise API consumption. Beyond the two leading labs, additional model providers, orchestration platforms, application hosting services, retrieval infrastructure, and security tooling contribute meaningfully to the overall figure. Netting out consumer revenues and accounting for these surrounding software layers, agent stack spending in 2026 could be in the range of $80 billion to $100 billion.

New agent-generated value

Future agentic revenues arrive from a conceptually new layer formed by the machine economy. This will be a byproduct of the collapse of thinking power, bringing previously uneconomical workflows within reach. Put another way, machine intelligence will eventually become cheap and reliable enough to do work that humans previously skipped, such as the tasks deemed too small, frequent, or expensive to justify the opportunity cost.

This follows a pattern Amias Gerety, partner and head of US at QED Investors, sees across most technology waves. According to Gerety, "Technology almost always moves in three phases. The first is cost efficiency. The second is consumption increase—as the price goes down, you consume more. And the third, which is hardest to predict but ultimately the biggest, is completely new forms of economic activity."[14]

New agentic value unlocked by falling machine intelligence costs



Source: PitchBook

These could be tasks where agents continuously tweak interfaces for individual users based on hundreds of tiny signals, dispute minuscule billing errors that only become meaningful in aggregate, or maintain every internal record by closing stale tickets, archiving old files, and updating outdated tags the moment they go stale. A key component of this new agentic market is the interactions between machines without human involvement. With proper infrastructure, agents will be able to buy and sell tiny task level units of capability from each other.

Conceptual examples of machine-to-machine interactions in a new agent-only market

Scenario	Agent action	Counterparty	Format of exchange
Real-time traffic or event data purchase	Agent buys live traffic or incident data before rerouting or repricing logistics decisions	Connected vehicle network or roadside sensor operator	Pays per capture or per data pull
On-demand building telemetry	Building agent checks carbon dioxide, particulate matter, or occupancy data before adjusting HVAC or scheduling	Third-party telemetry service	Pays per reading or per minute of access
On-demand asset inspection	Agent hires a nearby drone or sensor to inspect an asset before triggering a maintenance or reorder decision	Autonomous drone operator or Internet of Things sensor network	Pays per inspection completed
Per-task agent delegation	Agent identifies a task outside its domain and delegates it to a specialized agent	Third-party domain-specific agent	Pays per completed task output
Microcredit line leasing	Agent borrows short-term credit to cover a temporary cash flow gap	Liquidity agent that fractionalizes a larger credit facility into microcredit slots allocatable to counterparties	Pays fees and interest for the duration the credit is held

Source: PitchBook

Each workflow generates its own demand, with agents becoming customers of other agents and creating downstream consumption of data, compute, and services. This compounds into a new layer of machine-native spending, achieved when advanced reasoning is accessible to all enterprises, not just hyperscalers. Intelligence is moving in this direction. Data from Epoch shows that the price to reach fixed LLM benchmark scores has fallen by 9x to 900x per year, with a median decline of about 50x. Meanwhile, across AI class chips released between 2012 and 2025, the compute performance available per dollar has improved by approximately 40% per year, roughly translating to a 2x gain every two to three years.[15]

However, this market is incredibly difficult to size today. Many of these agent interactions do not exist yet, and the number of possible agents and workflows has no obvious limit. One early attempt to measure the economic value already being generated by agents comes from Forsy, a startup that is building the learning loop for agents operating in real-world and cyber-physical systems. As of July 2026, the company estimates $36 billion in annual global agent GDP on a run-rate basis—a metric it defines as the net economic value added by deployed agents, adjusted for overlap with human work, software, and agent cost of goods sold.[16]

Forsy's agent GDP is not a measure of total AI spending or total revenue from AI companies. Instead, it focuses on outputs where an agent was the primary driver of the result. This comes from four key components, adjusted for overlap:

1. **Agent-assisted work value:** Output where an agent improves a human's speed or quality, but the human owns the workflow.

2. **Agent-generated revenue:** Revenue where an agent completes the commercially decisive step.

3. **Agent service revenue:** Income from deliverables such as research reports, support resolutions, or operations outputs.

4. **Agent asset revenue:** Reusable digital assets such as code, templates, or knowledge bases that retain value beyond a single task.

Their calculation begins with a country's economic base, reduced to work that is digitally addressable, then to tasks that agents can perform, and finally to work that is already being deployed by agents. Deployment is assessed through signals such as enterprise adoption, API usage, public tooling adoption, and other deployment indicators. A final step determines how much of the output should be credited to the agent versus the human or software around it, with higher attribution when the agent completes more of the workflow end-to-end.

Top 10 countries by agent GDP

Country	Agent GDP (per month)	Agent share of digital work	Median agent stack cost (per month)	Median revenue from agent-enabled work (per month)	Economic value generated per dollar of agent stack cost
US	$1.4B	1.0%	$160	$1,190	4.3x
China	$540.0M	0.7%	$50	$640	4.3x
UK	$198.0M	0.8%	$132	$805	3.7x
India	$182.0M	0.7%	$24	$325	4.3x
Germany	$172.0M	0.7%	$128	$745	3.6x
France	$131.0M	0.6%	$112	$675	3.4x
Japan	$120.0M	0.5%	$108	$620	3.2x
South Korea	$76.0M	0.6%	$98	$585	3.3x
Canada	$51.0M	0.5%	$98	$605	2.9x
Australia	$37.0M	0.5%	$92	$560	2.8x

Source: Forsy.ai • Geography: Global • As of July 2026

An important takeaway from Forsy's data is that the early machine economy is already producing measurable output. The median agent-enabled operator is generating about $90 per month in revenue directly attributable to agent work, and agents are producing 1.5 times the output of a human doing the same work without one. Even so, global agent GDP of $3 billion per month across 200 countries is still small.[17] As the enabling infrastructure matures and intelligence becomes cheaper and more capable, the market should scale materially.

In this section, we explore evidence of how machines are rapidly ingesting knowledge work and becoming dominant actors across the web.

Machine economy signals

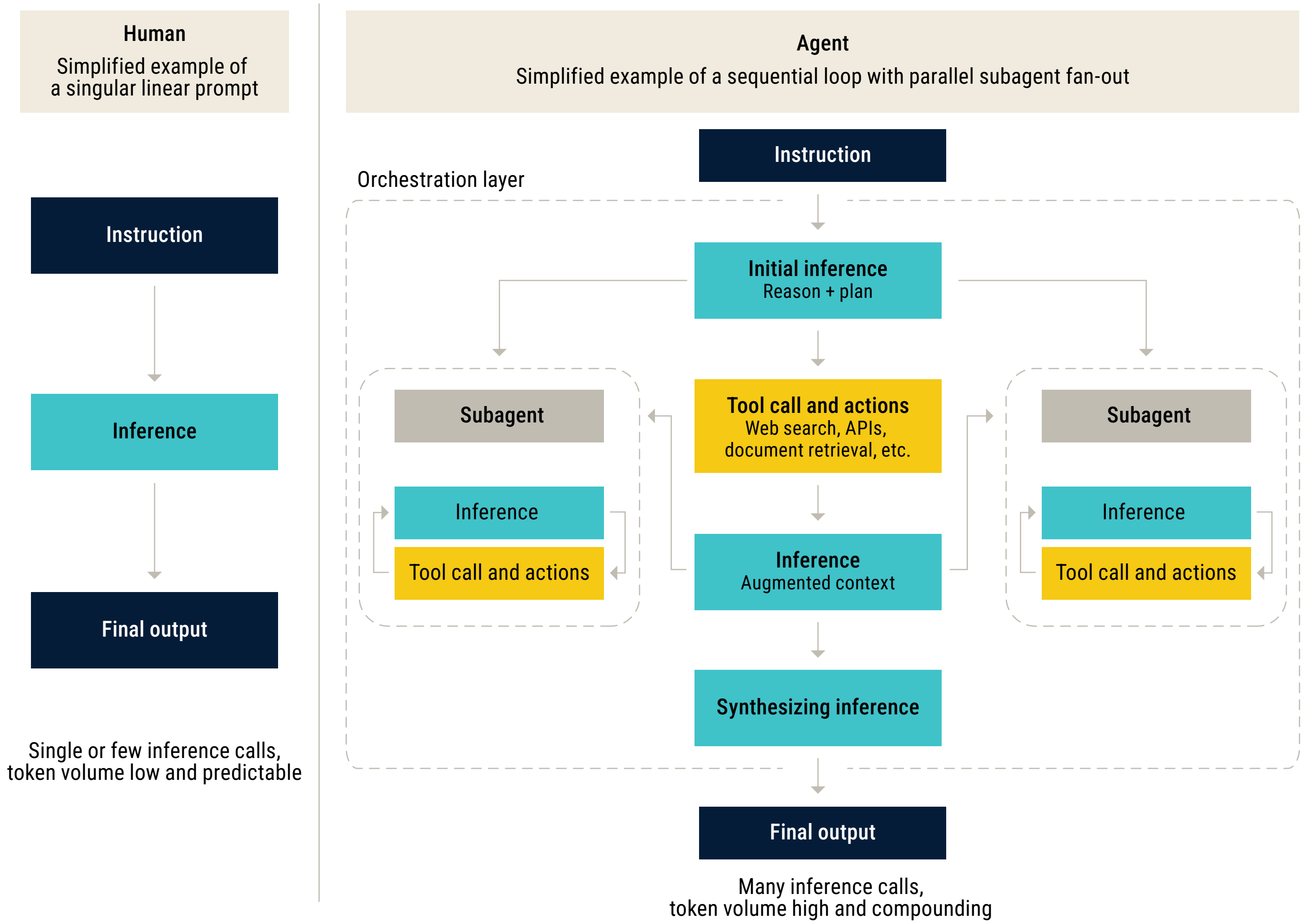
Strong evidence already exists that machines are creating an increasing share of value across the internet and software stack. Token volumes are compounding at rates human usage alone cannot explain. Bots now account for most HTTP requests on the web. APIs are being repriced for machine-scale consumption, and the protocols that let agents plug into external software are proliferating faster than any prior developer standard. We examine each of these signals below.

Model API throughput

API throughput of LLM providers can help capture how much work agents are doing. It is essentially the amount of model work performed via an API, typically measured as the volume of tokens processed via programmatic requests over a given period.

Observing this metric helps explain the explosive growth in the number of tokens processed, as agents interact with APIs in fundamentally different ways than humans do. Every LLM-powered action, whether a decision, summary, code commitment, or transaction, begins with an inference call. When a human uses AI, traffic is linear—a single prompt often yields a single response. But for an AI agent, it can trigger an exponential fan-out effect, where a single instruction branches into many background inferences as the agent plans, invokes tools, and checks its own outputs.

Example of agent fan-out effect



Source: PitchBook

The ability of agents to exponentially multiply calls at this scale and speed is pushing broader cognitive workloads onto machines. It is one of the contributing factors to the immense growth in tokens processed. Data from OpenRouter and a16z indicate that the average prompt length nearly quadrupled between early 2024 and late 2025, reflecting more complex, multistep instructions being given to models.[18] 2026 trends show this trajectory continuing. For example, Google's monthly token processing rate was 9.7 trillion in May 2024, 480 trillion in May 2025, and 3.2 quadrillion tokens in May 2026.[19]

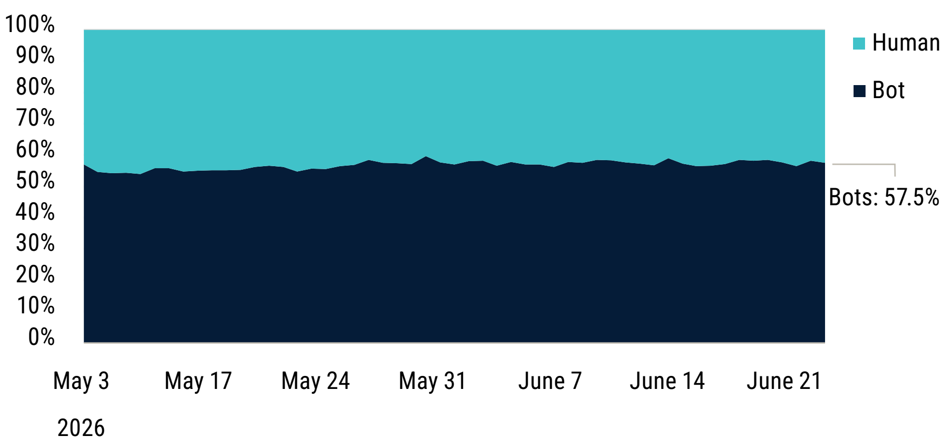
Narrowing to the model level, provider disclosures show how much of this growth is driven by machine-initiated API consumption. OpenAI's APIs processed 15 billion tokens per minute at the end of March 2026, with enterprises now accounting for more than 40% of revenue and projected to match consumer sales by the end of 2026.[20] Alphabet reported in Q1 2026 that its Gemini models were processing more than 16 billion tokens per minute via direct API usage, up 60% QoQ. It also sees usage concentrating among heavy enterprise workloads, with 330 Google Cloud customers each processing more than 1 trillion tokens and 35 reaching the 10 trillion token milestone over the past year.[21]

The throughput numbers seen today suggest we are inevitably building toward a separate agent-led economy. Tens of billions of tokens are processed through model APIs every minute, with a growing share driven by enterprise workloads and agents. But adoption is still early because most agents today operate within supervised, narrowly scoped systems. What we are measuring today is likely just the new baseline, and the bulk of machine-initiated inference has yet to come.

Internet traffic

Signals in web-crawling data and the composition of internet traffic are beginning to show that machines may be becoming the internet's dominant actors in terms of volume. While Cloudflare CEO Matthew Prince predicted in March 2026 that bot traffic would exceed human traffic on the web by the end of 2027, that crossover has already arrived.[22, 23] Bots now surpass humans in HTTP requests to HTML content. As of June 24, bots accounted for 57.5% of these requests, compared with 42.5% for humans.[24]

Share of HTTP requests by user type



Source: Cloudflare Radar • Geography: Global • As of June 24, 2026

However, “bots” is a broad category. It includes legacy automated traffic, such as search engine crawlers, uptime monitors, and security scanners, as well as malicious bots used for credential stuffing, scraping, and fraud. It also includes AI training crawlers that bulk-fetch web pages for model development, which consume the web passively rather than acting on behalf of users.

Despite the volume coming from other bot categories, AI remains the fastest-growing driver of automated traffic. Training crawlers make up 67.5% of AI-driven traffic observed by HUMAN Security, but that share declined through 2025 as agentic categories accelerated. Scraper traffic grew 597% year over year, while agentic AI traffic grew 7,851%, albeit from a low base.[25] Browser-based agents that spin up full browser environments, load pages, click links, and fill forms now account for around 71% of HUMAN's observed agentic traffic, providing early evidence of agent-led activity across the web well before fully machine-native interfaces have been developed.[26]

Third-party APIs

After inference and reasoning, agents reach into the full stack of third-party services to execute tasks. External API volumes capture this downstream activity, such as retrieving data sources, initiating payments, or triggering software workflows. Agents do this through APIs, sending HTTP requests to endpoints that expose discrete functions. Because agents cannot navigate human-centric interfaces, this request-and-response cycle has become their default mechanism for interacting with external software, leaving measurable traces of autonomous activity across the web.

The effects of this shift have been apparent, as 65% of organizations generate revenue from their APIs.[27] Salesforce processed almost 1 trillion API calls in only the first quarter of 2026.[28] Alpaca, a brokerage infrastructure API provider, saw monthly API usage growth accelerate from single digits in Q4 2025 to roughly 30% in Q1 2026, which it attributed to AI agent-driven market participation.[29] The shift is also showing up in how developers build. About 25% now design APIs with agents as the primary end consumer, and 51% cite unauthorized agent access as their top security concern.[30]

Platforms are also repricing in response. Many are replacing flat-rate access with metered pricing calibrated to consumption types, with the goal of preventing infrastructure costs from ballooning while capturing revenue from high-volume automated usage. Xero moved to usage-based tiers in March 2026, charging on data volume and connections rather than a revenue-share model that agents could exploit by pulling large data volumes without generating proportional revenue.[31] In Q2 2026, X updated its API pricing, making it cheaper for users to read their own account data but 1,900% more expensive to post links to its platform via its API.[32] The changes reflect how APIs are being restructured for machine volumes.

Connective protocols

2026 has seen a sharp rise in the connective tissue that lets agents access software systems. Companies are increasingly launching their own MCP servers and CLI integrations—two complementary layers that enable agents to either understand and use external systems via structured tool access or operate directly on a user's machine. Both are utilized by AI agents depending on the use case.

Development around these connective protocols has accelerated as machine activity on the web and consumption of digital services have grown. Our query of MCP servers on June 18, 2026, shows that there are now 12,594 unique servers from when Anthropic introduced MCP in November 2024.[33] The growth rate is rising fast; there were 9,652 distinct servers roughly one month prior, on May 24, 2026.[34] However, these public counts likely understate actual deployment, as many servers are not indexed or publicly listed. The actual number of MCP servers running in production is likely substantially higher. For example, our query of the GitHub Search API returned over 18,000 publicly shared MCP server codebases.

CLI utilization is also growing in tandem. The interface is resurging because its text-based format is machine-friendly and avoids inflating context windows. Stripe reports that 70% of its CLI requests for API resources now come from agents and that agent traffic to its documentation grew more than 10x in 2025.[35] At Checkly, 50% of CLI users are now agents, and the company predicts an even greater share by the end of 2026.[36] Inside Uber, 84% of the company's developers are now agentic coding users, using either CLI-based agents or AI tools embedded directly in their coding software.[37]

Because agents are emerging as a primary user class, companies are reshaping their products for them. In 2026, many companies have already launched CLIs geared toward agents, including Visa, Ramp, Mercury, ElevenLabs, Stripe, Coinbase, MoonPay, and DoorDash. There is a flywheel effect to note here: As more companies launch agent-native CLIs, agents gain broader access to execute work, driving further adoption. Companies will then build more agent-first infrastructure, accelerating the cycle.

Why agentic economic actors are not priced in yet

In this section, we discuss the scope of autonomy for agents today and the factors that limit their ability to become economic actors.

Clear evidence indicates an agent economy is forming, with the potential to scale to unprecedented size. However, agents are not yet priced as true economic actors because their autonomy remains constrained. To matter economically, agents must be more than advanced assistants executing narrow tasks within session-level workflows and predefined tools. Autonomous agents must operate across tasks, adapt to changing contexts, carry work forward over time, and commit resources without waiting for approval.

Key limitations of agent autonomy

Dimension	How to think of it	How it works today	How autonomy breaks
Scope	What agents are used for	Agents can handle parts of a workflow, such as summarizing data or screening vendors.	Agents cannot own the full process or complete cross-system tasks. For example, finding, evaluating, and procuring a third-party dataset without human help.
Controls	Which tools can agents touch	Agents operate with limited, scoped access, and identity is usually tied to the human or organization running them.	Agents cannot reliably prove who they are or take state-changing actions on their own, such as moving money, signing up for services, or updating systems of record.
Supervision	Who has the final say	Humans review and approve key steps, especially those with financial or operational impact.	Agents cannot bind the business to a transaction, such as completing a purchase or entering into a vendor agreement, without human approval.
Longevity	How long can agents operate	Memory is limited or fragmented, and sessions often carry little context forward.	Agents cannot persist as durable economic actors over time, such as independently maintaining context, relationships, and history.

Source: PitchBook

In other words, today's agents are strong at executing within predefined boxes but weak at independently owning workflows. Underlying these limitations is also another fundamental issue: Agents are not always fully capable of being trusted with broader autonomy across most domains. Today's agents perform well on computational and coding tasks, where outputs are verifiable, and environments are structured.

However, real-world workflows introduce novelty and uncertainty that most agents were not trained to handle, such as edge cases, ambiguous instructions, and multisystem dependencies that fall outside their training distribution. Ray Ren, co-founder of Forsy, describes this problem as a major gap for agents today. According to Ren, "There is a huge gap between the environments agents learn in and the environments they are deployed into. In controlled settings, agents can improve through clear rewards, repeatable trials, and dense feedback. Real-world work is different because outcomes are sparse or delayed, conditions keep changing, and the same workflow often cannot be reset and replayed."[38]

To provide a snapshot of where agent autonomy stands today, we scored 40 deployed agent products on two axes using publicly available evidence, such as product documentation, demos, and customer case studies. Each axis includes a score from 1 to 5 that reflects an agent's level of autonomy and scope. These scores captured how agents are actually used today, rather than the technical ceiling they could reach if granted full permissions.

Autonomy: considers whether a human must approve each action, what triggers the agent's work, whether it can plan a sequence of steps, what actions it can take on its own, whether it retains memory across sessions, and how long it can run unattended. We assigned autonomy scores based on the following:

1. Acts only when prompted, with no independent planning or initiative.

2. Uses tools and acts within a single prompted session, with narrow permissions and little memory across sessions.

3. Pursues multistep goals across tools but needs human approval for high-impact actions and usually starts fresh each run.

4. Runs for extended periods under set guardrails, replans from results, and acts without step-by-step prompts, while humans mainly monitor and handle exceptions.

5. Maintains its own identity and memory over time and initiates actions or transactions under broad delegated authority, with humans setting the goals and rules rather than approving individual steps.

Scope: measures how much of a task the agent owns from start to finish, how many systems it works across, and the kinds of decisions it is trusted to make. We assigned scope scores based on the following:

1. A single task, such as drafting a message or answering a question.

2. A few linked steps within a narrow slice of work, such as researching and then summarizing.

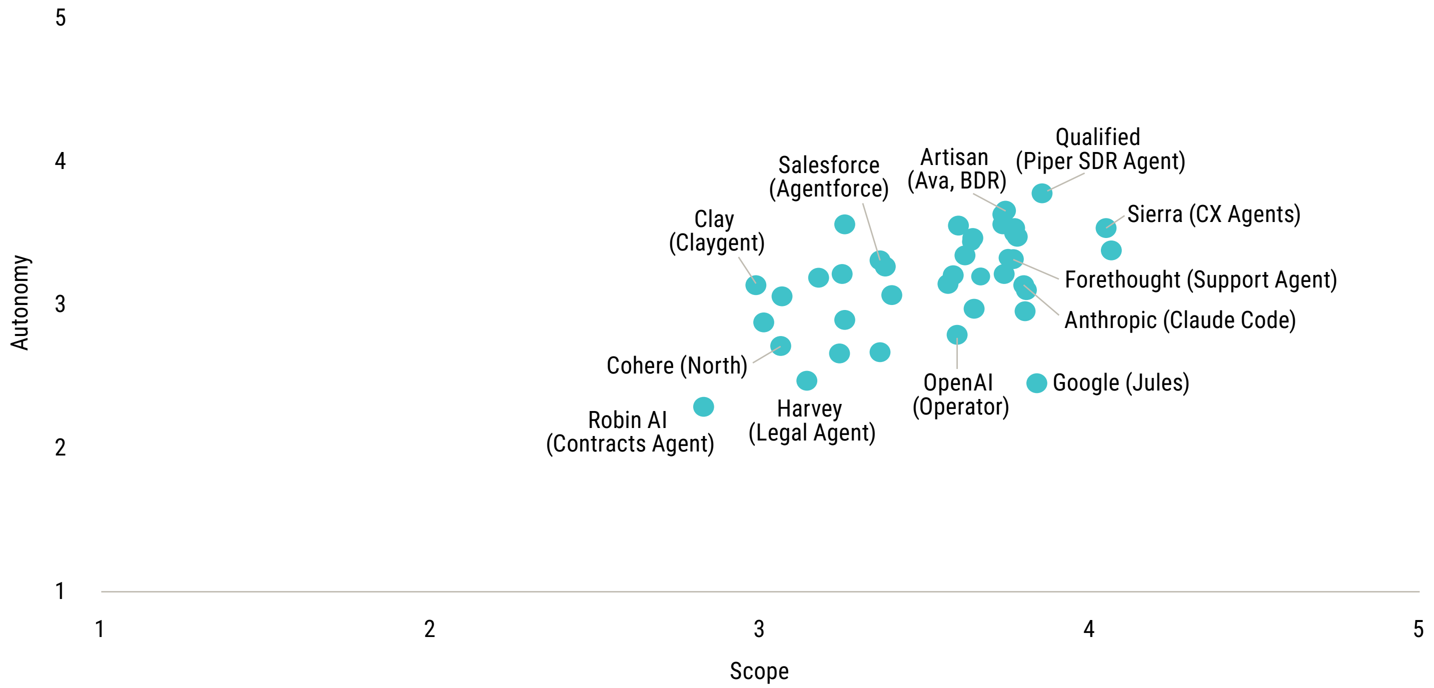
3. One complete stage of a larger workflow, where a person or a later step still carries the result forward.

4. A full outcome within a single narrow domain, completed without a handoff and often within set limits.

5. A broad business outcome that coordinates several workflows or functions across multiple systems. For example, managing replenishment across inventory, ordering, and supplier systems.

New agent-generated value and economic activity become plausible when agents reach the upper end of the autonomy scale. We found that our scores clustered primarily in the middle of the grid, with autonomy averaging 3.2 and scope averaging 3.5. This highlights how, in today's deployments, agents operate as supervised assistants that own a portion of a workflow rather than as autonomous actors that run independently.

Out-of-the-box AI agents: Scope versus autonomy



Source: PitchBook • Geography: Global • As of June 30, 2026

There are several reasons for this. First, enterprises lack the infrastructure to safely grant broader autonomy, including reliable system access, observability, and rollback. Second, risk remains a binding constraint, especially around authorization and liability. Third, agents still lack the ability to transact independently. Without native ways to exchange value, they remain limited to tightly controlled environments.

Closing the distance between an agent that completes a single task and one that acts across systems, counterparties, and money is where opportunities are rapidly emerging. Below, we isolate where these gaps occur in an agent's workflow.

In this section, we walk through an end-to-end workflow that demonstrates where autonomy breaks down due to limitations in machine-first infrastructure.

POV: You are an AI agent

Agents operate in a chain of workflows, which are essentially control loops that turn an instruction into a sequence of reasoning steps, tool calls, and checks. However, this flow often breaks down and requires a human to intervene because the necessary infrastructure has not been fully developed. To illustrate this, we walk through a step-by-step process of an agent executing a simple task.

Example scenario

A human on a forecasting team tasks an agent with sourcing and acquiring an external dataset to improve the accuracy of an internal demand model. The task specifically involves sourcing and acquiring anonymized transaction data covering order volume, average basket size, and regional purchase frequency over the past 90 days.

Step 1: Instruction intake and initial inference

The agent receives the task in plain text. Inference occurs in an LLM as the agent interprets the request, identifies the objective, and reasons through what the task requires. Afterward, the model determines if any additional context is required.

Instruction intake and initial inference

Source: PitchBook

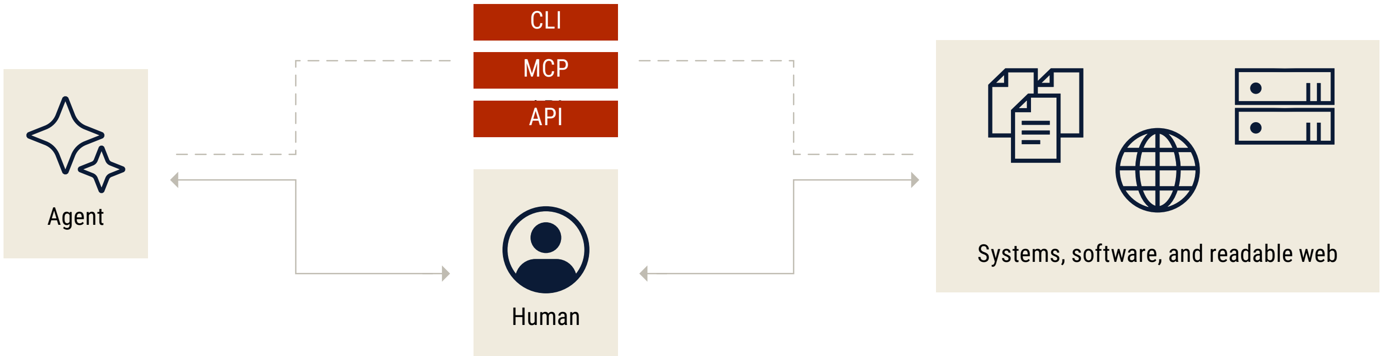
Step 2: Additional inference for context augmentation

In this case, the agent determines it needs additional context surrounding the internal model's data requirements, budget limits, and prior vendor evaluations. It also needs to identify additional vendor options and how to access them. An orchestration layer triggers the retrieval-augmented generation process, pulling relevant information into the working context for another inference pass.

If CLIs were integrated or MCP servers were connected, the runtime could cleanly retrieve structured data by calling internal systems. These could include a document management system, a customer relationship manager with prior vendor records, or an internal data catalog.

For this example, no prior integrations were available. As a result, the agent must fall back to asking the human for missing details and searching external vendor pages for clues about coverage, pricing, and access. Because the web is not optimized for machines, the agent must attempt to parse unstructured HTML and page text, which lacks the structured data an API would provide.

Additional inference for context augmentation

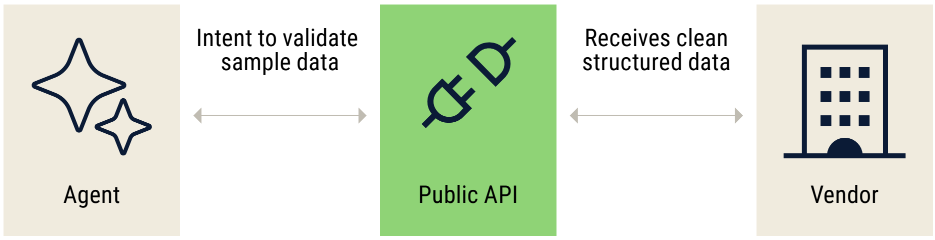


Source: PitchBook

Step 3: Vendor identification and validation

After additional inference occurs with augmented context, the agent selects a potentially suitable vendor it discovered from the previous step. Luckily, this vendor exposes both a public and authenticated API. The agent can therefore call the public endpoint, which requires no credentials and is set up to allow prospective buyers to evaluate the data before committing. After pulling a sample, the agent verifies that the dataset is sufficient.

Vendor identification and validation



Source: PitchBook

Step 4: Dataset retrieval

Because the sample dataset passed validation, the agent proceeds to retrieve the full dataset. This time, it must call the authenticated API endpoint.

However, the agent cannot proceed on its own. No API key has been preloaded into the orchestration layer for this vendor, and the agent has no mechanism to acquire one on its own. This is because the vendor's credentialing process is human-facing, meaning an account must be created through a web browser, a user must be authenticated, an API key must be issued, and money must be fronted. As a result, a human re-enters the flow, manually acquiring the API key and provisioning it into the orchestration layer.

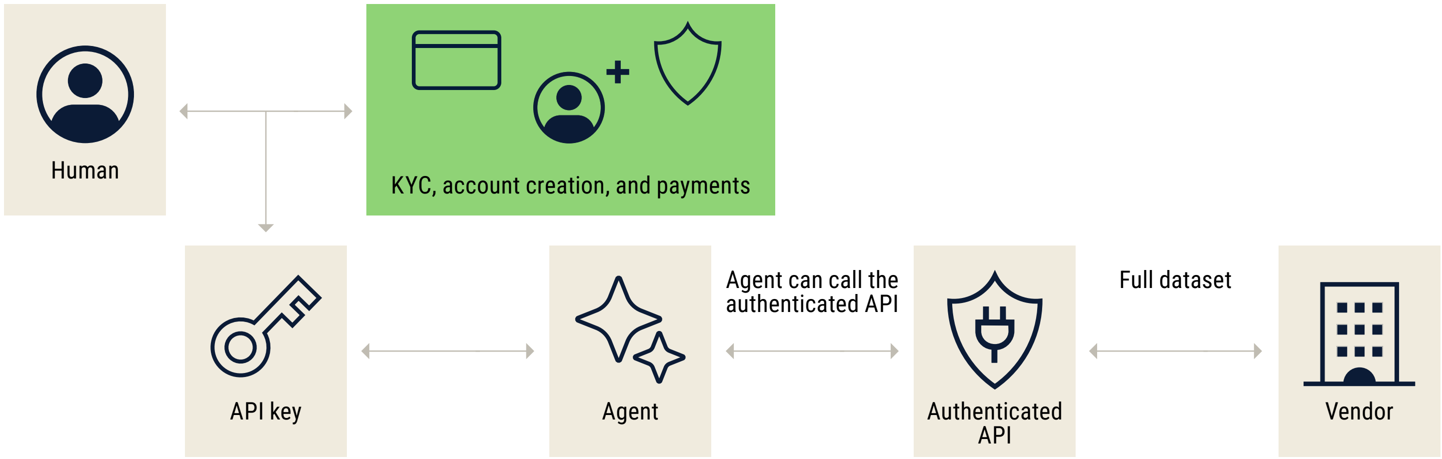
Dataset retrieval

Source: PitchBook

Step 5: Task completion

With credentials provisioned, the agent calls the authenticated API endpoint, draws down on prefunded credits, and retrieves the full dataset. For the purposes of this example, the task is complete. However, other scenarios could involve the agent spawning subagents to clean the data, cross-referencing it with other datasets, or pulling data in real time to keep the model current. These subagents would similarly run their own reasoning loops and may require their own credentials and ability to make payments.

Task completion



Source: PitchBook

Scenario analysis

The agent successfully completed its task in our example, but only because a human filled in the gaps at key moments. As shown above, the agent failed to remain fully autonomous in steps two and four, which are essential to the workflow.

Both structural and variable issues prevented the agent from acting autonomously. Any agent running a similar task is likely to encounter the same structural roadblocks with authentication and payments. One reason is that authenticated APIs, the interfaces through which agents access and transact across external services, require credentials. However, this challenge will persist for agents until there are more widely accepted methods for authorizing agent identity. In addition, without wallets or delegated payment authority, the agent cannot settle costs at the point of transaction without human intervention.

Variable issues also shape whether an agent can discover or use what it needs in the first place. In our scenario, the agent was fortunate to find a suitable vendor that also exposed an API. Many vendors still do not present information in machine-readable or discoverable formats, and bot defenses may block automated access altogether. This creates a supply-side gap in the machine economy and reinforces the need for machine-first infrastructure to support an agent economy.

Part I concluding takeaways

The machine economy is a multitrillion-dollar market emerging as agents discover, transact, and generate value with little to no human involvement. Its value comes from three layers: existing software and labor spending shifting to agents, the infrastructure agents consume to run, and new value created by an agent-only market that did not exist before. Evidence from token throughput, web traffic, API consumption, and the rapid rise of machine-facing protocols like MCP and CLI suggests this market is already forming.

For companies and developers, this means that building for agent traffic will become nonnegotiable. Success will increasingly depend on:

- Making products machine-readable and APIs easy to use through service directories, standardized data formats, and command-line access.
- Designing pricing, rate limits, and observability specifically for agent traffic so consumption is predictable and billable.
- Treating tools and data as callable services that can be metered per call, per outcome, or per agent-execution rather than through human-seat licenses.

As agents gain access to more software and infrastructure, new revenue models will emerge around how often services are used and what results they produce. Some services will be charged by the call, others by the outcome, and others by the amount of data or processing consumed. Over time, agents may buy services from other agents, and the companies ready for that shift will be best positioned to benefit.

But a fully realized machine economy is still far off, and we estimate that only about 1% of today's AI-exposed work is handled by agents. Today's agents are useful, but they remain constrained by limited scope, weak controls, human supervision, and short-lived context, which prevents them from acting as true economic actors. This gap is most obvious in today's agentic workflows, where autonomy typically breaks down at three structural points: when information is not machine-readable or discoverable, when an agent lacks the credentials or tools required to act, and when payments require human intervention.

Solving the payments layer is especially difficult. For these systems to work, agents must be trusted to hold funds, initiate transactions, and operate within defined spending limits. Even then, questions remain around identity verification, intent, real-time micropayments, and how to reverse transactions when something goes wrong.

In Part II, we will examine the emerging solutions designed to close these gaps, how they work, and which approaches are likely to win.

References

1: "Agent Economy," Forsy, n.d., accessed July 12, 2026.

2: "I/O 2026: Welcome to the Agentic Gemini Era," Google, Sunder Pichai, May 19, 2026.

3: "OpenAI Raises $122 Billion to Accelerate the Next Phase of AI," OpenAI, March 31, 2026.

4: "Alphabet Q1 2026 Earnings Call," PitchBook, April 29, 2026.

5: "Traffic," Cloudflare, June 24, 2026.

6: "Stripe Projects Adds New Agent Integrations, More Providers, and Custom Developer Controls," Stripe, Rami Banna, June 11, 2026.

7: "Salesforce Delivers Record First Quarter Fiscal 2027 Results," Salesforce, May 27, 2026.

8: "State of the API Report," Postman, 2025.

9: "Ramp AI Index," Ramp, n.d., accessed July 16, 2026.

10: Ibid.

11: "2025: The State of Generative AI in the Enterprise," Menlo Ventures, Tim Tully, et al., December 9, 2025.

12: "Anthropic Raises $65B in Series H Funding at $965B Post-Money Valuation," Anthropic, May 28, 2026.

13: "OpenAI Raises $122 Billion to Accelerate the Next Phase of AI," OpenAI, March 31, 2026.

14: Amias Gerety, partner and head of US at QED Investors, phone interview by Rudy Yang, May 19, 2026.

15: "Trends in Artificial Intelligence," Epoch AI, February 5, 2026.

16: "Agent Economy," Forsy, n.d., accessed July 12, 2026.

17: Ibid.

18: "State of AI," OpenRouter, Malika Aubakirova, et al., December 2025.

19: "I/O 2026: Welcome to the Agentic Gemini Era," Google, Sundar Pichai, May 19, 2026.

20: "OpenAI Raises $122 Billion to Accelerate the Next Phase of AI," OpenAI, March 31, 2026.

21: "Alphabet Q1 2026 Earnings Call," PitchBook, April 29, 2026.

22: "Bots Now Outnumber Humans Online and the Internet Was Never Built for This," Forbes, Josipa Majic Predin, June 4, 2026.

23: "Welp, that happened faster than I predicted…" X, Matthew Prince (@eastdakota), June 3, 2026.

24: "Traffic," Cloudflare, June 24, 2026.

25: "The 2026 State of AI Traffic & Cyberthreat Benchmark Report," HUMAN Security, n.d., accessed July 12, 2026.

26: "State of Agentic Traffic – April 2026: Agentic Browsers Generate Nearly Three-Quarters of Agentic Traffic," HUMAN Security, Aviad Kaiserman, May 4, 2026.

27: "State of the API Report," Postman, 2025.

28: "Salesforce Delivers Record First Quarter Fiscal 2027 Results," Salesforce, May 27, 2026.

29: "Alpaca Reports Sharp Growth in API Trading as AI Reshapes Market Access," Alpaca, Luke West, April 23, 2026.

30: "State of the API Report," Postman, 2025.

31: "Frequently Asked Questions," Xero, n.d., accessed July 12, 2026.

32: "X API Pricing Update," X, April 16, 2026.

33: "Official MCP Registry," Model Context Protocol, n.d., accessed July 12, 2026.

34: "MCP Adoption Statistics 2026: Model Context Protocol," Digital Applied, April 20, 2026.

35: "Stripe Projects Adds New Agent Integrations, More Providers, and Custom Developer Controls," Stripe, Rami Banna, June 11, 2026.

36: "One CLI, Two Audiences: How We Built for Agents and Humans," Checkly, Stefan Judis, March 26, 2026.

37: "How Uber Uses AI for Development: Inside Look," The Pragmatic Engineer, Gergely Orosz, March 10, 2026.

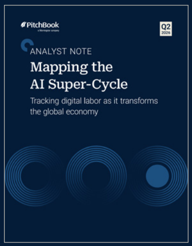
38: Ray Ren, co-founder of Forsy, phone interview by Rudy Yang, May 19, 2026.

PitchBook provides actionable insights across the global capital markets.

Additional research:

2026 Advanced Software Launch Report: Investing in the Great AI Margin Super-Cycle

Download the report here



Mapping the AI Super-Cycle

Download the report here

PitchBook Insights is an online compendium of in-depth data, news, analysis, and perspectives that shape the private capital markets.

PitchBook subscribers enjoy exclusive access to a comprehensive suite of private market insights, including proprietary research, news, data, tools, and more on the PitchBook Platform.

Nizar Tarhuni
Executive Vice President of Research and Market Intelligence

Paul Condra
Senior Director, Global Head of Private Markets Research

James Ulan
Director, Industry & Technology Research

Report created by:

Rudy Yang
Senior Research Analyst, Enterprise Fintech and Retail Fintech

Jenna O’Malley
Senior Graphic Designer

Learn more about PitchBook’s Institutional Research team.

Click here for PitchBook’s report methodologies.