

RESEARCH REPORT

The Agent Economy Gets Real: Revenue, Deflation, and the Discipline of Production AI

In eighteen months, AI agents went from conference demos to a revenue business: a handful of named agent products now generate roughly \$2.9 billion in combined annual recurring revenue, and venture funding for pure-play agentic AI hit a record \$6.42 billion in 2025. But the structure of the market has inverted — 144% more capital in early 2026 moved through 39% fewer rounds, into companies that already have customers, while new startup formation has frozen almost to zero. Chainfir Research maps the revenue proof, the inference-cost deflation engine underneath it, the deployment gap that separates experimenters from scalars, and where a profit-first investor should actually look.

Industry Reports · October 3, 2026 · Chainfir Research

Agents stop being demos: the revenue line has been crossed

The most important fact about AI agents in 2026 is no longer a capability benchmark — it is a revenue ledger. An April 2026 analysis by EigenCloud tallied roughly \$2.9 billion in combined annual recurring revenue across named agent products with verifiable pricing: Cursor, the AI code editor built by Anysphere, crossed \$2 billion in annualized revenue in February 2026, doubling in three months, after passing \$500 million ARR in June 2025 and \$1 billion in November 2025. Legal agent Harvey reached \$195 million ARR after tripling revenue during 2025. Sierra, the enterprise customer-experience agent platform founded by Bret Taylor and Clay Bavor, surpassed \$150 million ARR — having hit its first \$100 million in just seven quarters. Replit's agent scaled to approximately \$150 million, and Cognigy now processes more than one billion customer interactions a year.

The depth behind those numbers matters more than the headline. Cursor reports that 64% of the Fortune 500 uses its product, across more than 50,000 enterprises, with over 100 million lines of enterprise code written through the platform daily. These are not assistants waiting for prompts: Cursor orchestrates multi-step coding workflows, Harvey drafts settlement language on nine-figure cases, and Sierra's agents execute actions inside Fortune 500 customer operations. The agent economy's first revenue chapter is written in developer tools and customer service — the two categories where work is digital, measurable and abundant.

Named AI agent products: annualized recurring revenue (\$M)

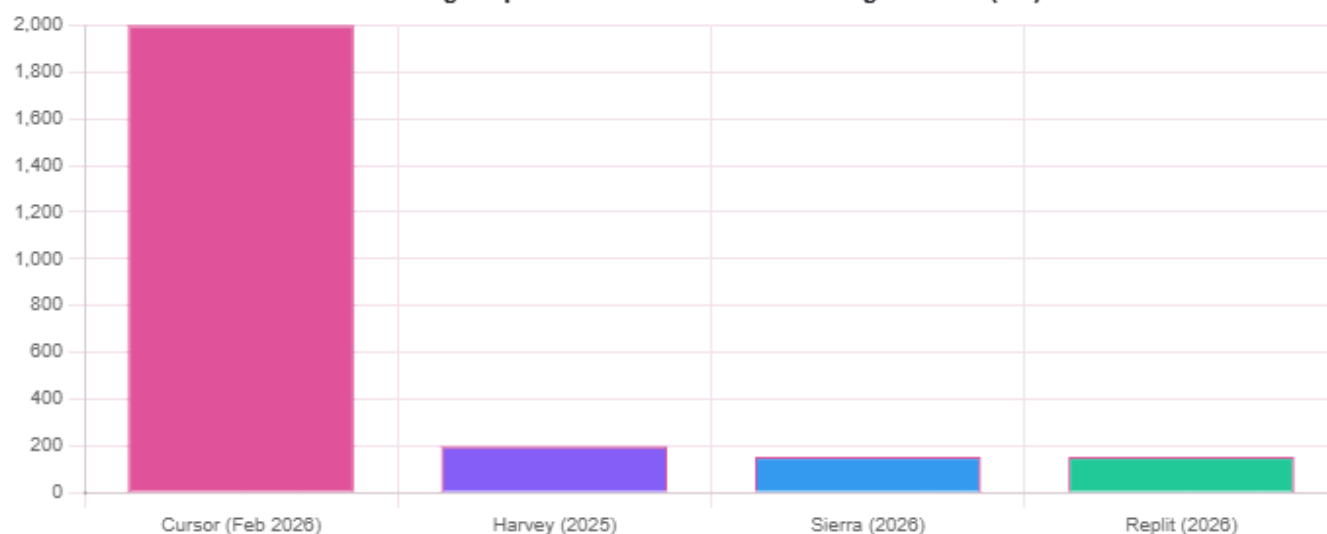


Chart 1: Verifiable ARR of leading agent products. Cursor doubled from \$1bn (Nov 2025) to \$2bn+ (Feb 2026); Harvey tripled during 2025 to \$195M; Sierra reached \$150M within eight quarters of launch. Named agents total roughly \$2.9bn combined ARR. Source: EigenCloud analysis (April 2026); company disclosures.

The deflation engine underneath

None of this economics works without the steepest cost curve in computing history. When GPT-4 launched in March 2023, frontier intelligence cost \$30 per million input tokens. By July 2024, GPT-4o mini delivered GPT-4-class quality for \$0.15 — a 200-fold decline in sixteen months, according to tokencost.app's AI Price Index, which tracks every major provider pricing change. Measured as the cheapest available GPT-4-class model, the price floor fell from roughly \$20 per million tokens in late 2022 to \$0.88 in Q4 2024 and about \$0.40 by Q4 2025 — a 50x decline in three years, according to Zhongtai Securities' industry deep-dive. Over the same period, PC compute costs fell about 3.4x and internet bandwidth about 2.6x. Intelligence is deflating an order of magnitude faster than any prior general-purpose input.

Two structural forces keep the curve falling. First, open-weight competition: Meta's Llama 3.1 and DeepSeek's V3 (launched December 2024 at \$0.14 per million input tokens, roughly 1% of GPT-4's launch price) repeatedly reset the price floor and forced every Western provider to reprice. Second, the frontier itself keeps cheapening: Anthropic cut Claude Opus by 67% overnight in late 2025, and Andreessen Horowitz estimates inference costs for equivalent capability have fallen roughly 10x per year since 2021. For agent builders the implication compounds: workloads uneconomical at last year's prices turn gross-margin positive this year, and agent gross margins — currently 40–70% versus 70–90% for traditional SaaS — expand mechanically as their largest cost line deflates.

OpenAI flagship input price per 1M tokens (\$), 2023–2026

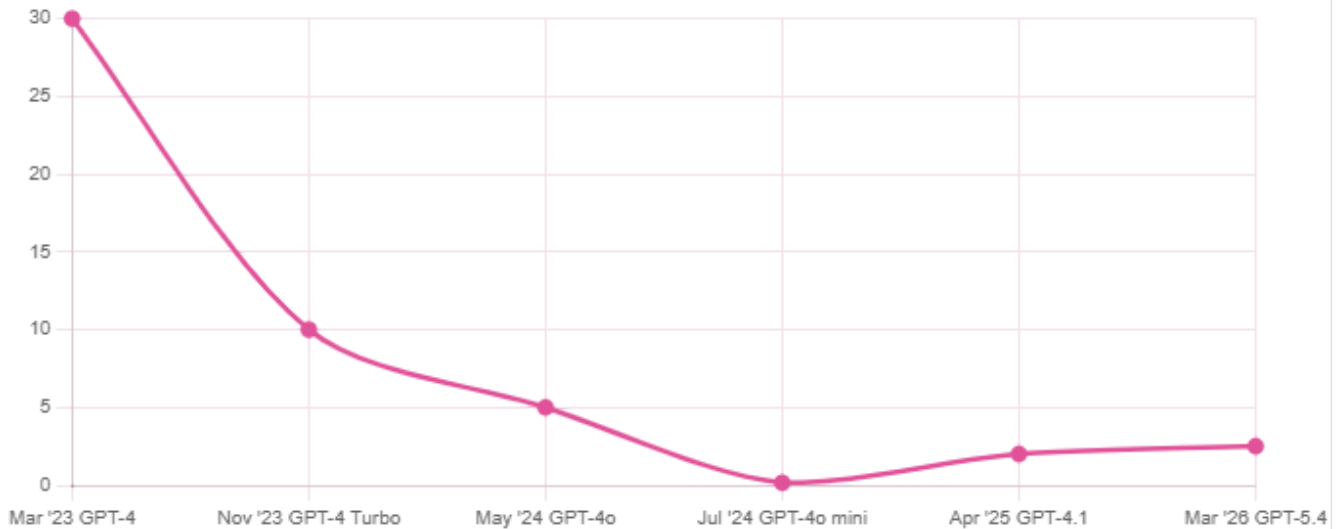


Chart 2: Frontier intelligence deflated from \$30 to \$0.15 per million tokens for GPT-4-class quality in 16 months (200x); the GPT-4-class price floor fell ~50x in three years to ~\$0.40 by Q4 2025. Source: tokencost.app AI Price Index (official API pricing pages); Zhongtai Securities industry report.

The capital signal: fewer, bigger, later

Venture capital has voted — and the vote is concentration. According to Tracxn sector data compiled by AgentMarketCap, pure-play agentic AI companies raised a record \$6.42 billion in 2025, capping a decade of \$24.2 billion in cumulative investment. But the 2026 pattern inverts the 2025 pattern: in January–April 2026, agentic AI companies raised \$2.66 billion across just 44 rounds, versus \$1.09 billion across 71 rounds in the same 2025 window — 144% more capital through 39% fewer transactions. Average round size roughly quadrupled from about \$15 million to \$60 million, and for the fifteen largest recent rounds it reached \$155 million, nearly double the \$82 million average of H1 2025. Perhaps most striking: 245 new agentic AI startups were founded in 2023; in the first quarter of 2026, Tracxn's live sector map recorded exactly one. The market is not cooling — it is sorting.

A New Market Pitch study of every disclosed pure-play agentic round from August 2025 to September 2026 — 39 financings, \$3.68 billion — shows the same compression: the top ten deals absorbed 69.1% of capital, Series B and later rounds took 82.4%, and 32 of 39 rounds were follow-ons returning to companies that had raised before. Geography is equally concentrated: North America produced 33 of 39 deals (\$2.66 billion), Europe six (\$1.02 billion). The outlier region is China, where KPMG records agentic AI funding of \$29.47 billion in 2025, up 39%, accelerating to \$9.45 billion in Q1 2026 alone — up 129.4% year on year — within a total Chinese AI PE/VC market of \$40.64 billion (+58.5%) across 2,336 deals. Capital everywhere is paying for production systems, not prototypes.

Agentic AI venture funding: capital up 144%, rounds down 39%

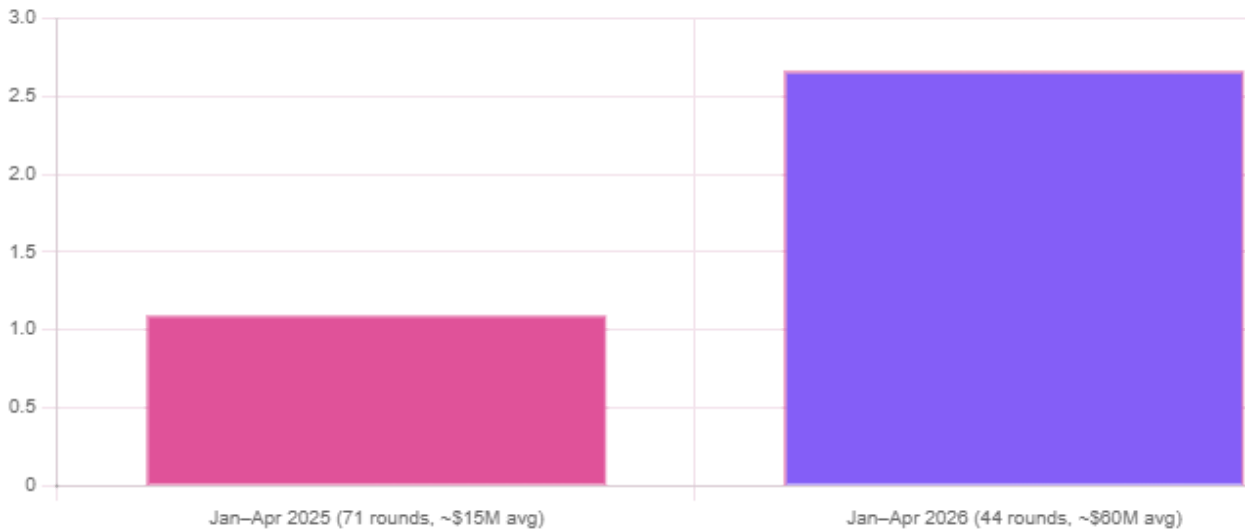


Chart 3: Same-window comparison of pure-play agentic AI funding. Capital rose 144% while round count fell 39%; average round size roughly quadrupled. Full-year 2025 was a record \$6.42bn. New startup formation collapsed from 245 in 2023 to one in Q1 2026. Source: Tracxn sector data via AgentMarketCap (April 2026).

Unit economics: outcome pricing replaces seat pricing

The business model innovation is as important as the technology. Intercom's Fin agent charges \$0.99 per resolved customer ticket — not per seat, not per message, per outcome — against a fully loaded human cost of \$15–25 per resolution: a 15–25x cost compression that EigenCloud calls "an economic inevitability." Klarna disclosed that its internal AI deployment replaced the work of 700 full-time customer-service equivalents, generating roughly \$60 million in annual savings — a narrative it carried into its IPO process. When software is priced against labor budgets rather than IT budgets, the addressable market expands by an order of magnitude: global customer service alone is a multi-hundred-billion-dollar labor line.

The efficiency shows up in the org chart as well. Anyosphere generates approximately \$13.7 million in ARR per employee, versus about \$290,000 at Salesforce — 47 times the revenue efficiency of the largest enterprise software company on earth, per EigenCloud's analysis. Harvey tripled revenue in a year without tripling headcount. This is what "digital labor" means in practice: the agent is not a tool the organization occasionally reaches for, but a process it now runs on — and that workflow ownership, not model sophistication, is what the best valuations are actually pricing.

The deployment gap is the real market map

Between intent and production lies the defining chasm of 2026. MuleSoft and Deloitte Digital's survey of 1,050 IT leaders found 93% plan to introduce autonomous agents within two years. McKinsey's November 2025 State of AI survey found 62% of organizations are experimenting with AI agents — but only 23% are scaling them in at least one business function, and fewer than 10% across multiple functions. Gartner, drawing on a poll of 3,412 organizations, predicts more than 40% of agentic AI projects will be cancelled by end of 2027, citing escalating costs, unclear ROI and insufficient risk controls — even as it forecasts 40% of enterprise applications will embed

task-specific agents by end of 2026, up from under 5% in 2025. Both things are true: adoption is near-universal, and most of it fails to scale.

The failure pattern is consistent across McKinsey, Gartner and Deloitte research — organizational, not technical: governance frameworks absent before deployment, no observability tooling to see what agents actually do, no baseline metrics captured before pilots (so ROI can never be proven), and no accountable business owner. Security concerns have delayed or limited AI initiatives for 51% of service leaders. The minority that succeeds shares four attributes: infrastructure investment before deployment, governance documentation first, baselines captured before pilots, and dedicated ownership with performance accountability. For investors, the deployment gap is not a reason for caution about the category — it is why durable value accrues to vertical specialists and to the governance, identity and observability infrastructure that closes it.

Where the value pools

Capital has already mapped the answer. In New Market Pitch's dataset, vertical AI agents — systems built for specific domains such as cybersecurity, healthcare, legal, finance and customer operations — produced 22 of 39 deals and \$3.04 billion, or 82.6% of all disclosed capital. Tracxn's longer window shows customer service and sales alone captured roughly 37% of all agentic funding from 2022 through 2025. Valuation discipline is reasserting itself too: Finro's Q1 2026 dataset across 214 AI agent companies shows an average multiple near 30x ARR, with the dispersion explained less by growth than by workflow position — the agent that owns a mission-critical process (Sierra at ~67x) outprices the one that rents attention. Market-size estimates vary by definition, but direction does not: Grand View Research puts enterprise-specific agentic AI at \$3.67 billion in 2025, reaching \$24.5 billion by 2030 (46.2% CAGR), while broader definitions size the total market at \$10.9 billion in 2026 growing to \$50.3 billion by 2030.

For profit-first investors, we are watching five things in this market:

- **Vertical agents with production revenue in regulated domains.** 82.6% of recent capital went to vertical specialists; legal, finance, healthcare and compliance agents monetize outcomes, not seats.
- **The governance and observability layer.** The 62%-to-23% deployment gap is an infrastructure purchase order: identity, authorization, audit and monitoring for agents is the picks-and-shovels position.
- **Outcome-priced businesses with proven unit economics.** Per-resolution and per-task pricing converts inference deflation directly into margin expansion.
- **Asia-Pacific's under-priced formation.** North America took 33 of 39 recent pure-play deals; outside China's \$29.5 billion agentic funding wave, APAC formation remains thin — a structural entry window for regional capital.
- **Compounders, not sprinters.** With average multiples re-rating toward 30x ARR and 40%+ of projects forecast to be cancelled by 2027, durable workflow ownership beats demo velocity.

Pure-play agentic AI funding by category, Aug 2025–Sep 2026 (\$bn)

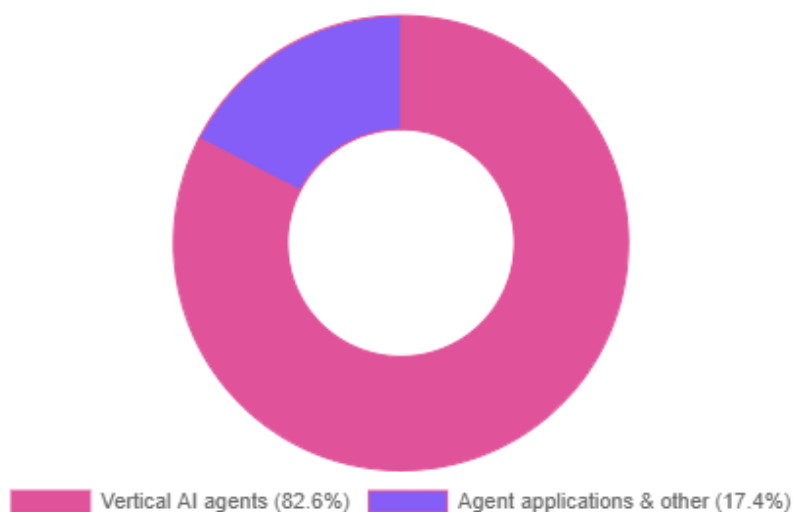


Chart 4: Of 39 disclosed pure-play agentic financings totaling \$3.68bn, vertical AI agents (domain-specific systems for legal, finance, cybersecurity, customer operations, etc.) took 22 deals and \$3.04bn — 82.6% of capital. Source: New Market Pitch agentic AI funding analysis (September 2026).

The profit-first reading

The agent economy's first chapter rhymes with every infrastructure cycle we have underwritten since 2018: the loudest capital goes to the frontier, but the durable returns accrue to the layer where capability meets a paying customer. A 200x collapse in the price of intelligence did not destroy value — it moved value from those who sell intelligence to those who apply it to work previously done by people. The companies now compounding — Cursor, Sierra, Harvey — share one trait with every durable technology business we study: their customers would notice immediately, in money, if the product disappeared.

That is precisely the filter our discipline applies. Chainfir Capital invests only in already-profitable companies — at least eight consecutive quarters of positive operating contribution — deploying roughly 90% of commitments from our own balance sheet, and integrating capital, talent and resources as one position. In a market where 62% of organizations are experimenting but only 23% are scaling, we are not interested in funding the experiment; we are interested in the companies on the right side of that ratio — already in production, already earning, already compounding — across Asia-Pacific's AI, Web3 and frontier-technology landscape.

The arithmetic is unusually clean this cycle. Intelligence deflates 10x a year; outcome pricing converts that deflation into margin; vertical workflow ownership converts margin into moats. The agent economy no longer needs to be believed. It needs to be underwritten.

"Two years ago the question was whether AI agents would ever make money. That debate is over — the revenue is real, unit economics proven, and intelligence costs keep falling 10x a year. The only question left is discipline: backing companies earning from production workflows, not demos. That filter has defined Chainfir since 2018, and this cycle rewards it more than any before." — Brian Tian, Founder & CEO, Chainfir Capital

