How Companies Are Pricing AI Features

Deep Research report · 150 sources

> The sources span two clear categories: pricing pages showing how ~40 companies actually charge for AI today, and infrastructure/commentary sources explaining why the market is converging on certain models. Together they show a market that started with simple bundling and is rapidly consolidating around usage-based credits, with billing infrastructure now a category of its own.

The Core Models in Use

1. Bundled free (AI baked into existing tiers, no separate charge)

The most common entry-level approach: give away a capped amount of AI as part of the base subscription price.

- Zoom includes meeting summaries, 20 AI queries/month, and AI note-taking in its free Basic tier, with a separate $20/month "ZoomMate" add-on for teams wanting more [31].
- Canva bundles an "AI allowance" (e.g., 200 Standard or 20 Premium AI uses) into every tier including Free, with an "AI Pass" add-on for extra usage on paid plans [41].
- Shopify checks AI features (Sidekick assistant, "millions of tokens," agentic commerce) across all four core tiers at no extra charge [48].
- xAI's Grok includes model access, real-time search, voice mode, and connectors in the Free tier, with higher tiers raising rate limits rather than charging separately for AI [143].
- Cursor's free Hobby tier includes limited Agent requests and Tab completions before usage-based billing kicks in at higher tiers [148].
- Lattice includes an "AI Agent" in all base products (Performance, Goals & OKRs, Engagement) at no extra cost [61].

2. Usage-metered / credit-based (the dominant emerging pattern)

By far the most repeated pattern across the pricing sources: a subscription fee plus a consumable "credits" currency that meters AI usage separately from seats.

- Figma allocates AI credits per seat type (3,000-4,250/month for Full seats, 500 for Dev/Collab seats) with a pay-as-you-go or subscription add-on for more [50].
- Airtable gives 500 to 25,000 AI credits per user/month depending on tier, with additional packs starting at $20/month for 10k credits [150].
- HubSpot uses "HubSpot Credits" as a cross-product usage currency for AI agents (Customer Agent, Data Agent), with an interactive calculator to estimate consumption [8].

- Lovable, Replit, Codeium, and Linear all price around monthly credit allowances blended with subscription tiers [76] [18] [43][137] .
- GitHub and LaunchDarkly gate free usage (2,000 Copilot completions/month; 5,000 AI runs/month) then bill overages, with LaunchDarkly charging $5 per 1,000 AI runs past the threshold [38] [59].
- Siena charges a flat $750/month platform fee plus $0.90 per automated ticket, a hybrid of platform access and per-action metering [1].
- Atlassian bundles Rovo AI credits per tier (25/70/150 credits per user/month) inside otherwise flat per-seat pricing [77].

3. Gated behind a paid tier

AI capability only unlocks once a customer moves up the plan ladder, rather than being metered or free everywhere.

- Zendesk's AI Agents, Knowledge Base, and Action Builder first appear at the $55/agent/month Suite Team tier; Admin Copilot and Intelligent Triage require even higher tiers, plus a separate $50/agent/month Copilot add-on [80][121] .
- Vanta bundles its core AI Agent (search, evidence checks, policy generation) free across tiers but gates more specific capabilities like AI-powered Questionnaire Automation as tiered/add-on features [9].

4. Paid add-on layered on top of a base plan

Distinct from gating: the base product is fully usable without AI, and AI is sold as an optional extra.

- Vanta's Questionnaire Automation and security-review features are explicit "Add-on" line items on top of otherwise-included tiers [9].
- Zoom's ZoomMate is a $20/month add-on for advanced agent capabilities beyond what's bundled free [31].
- Zendesk's Copilot add-on runs $50/agent/month on top of Suite plans [80].

5. Contact-sales / not disclosed

A meaningful share of enterprise vendors don't expose AI pricing publicly at all.

- Salesforce's pricing hub is entirely contact-sales, pointing instead to a "Flex Credit Pricing Calculator" for consumption-based products like Data 360 and Agentforce without disclosing actual rates [10][33] .
- Rippling shows no tiers or numbers, just a "request a quote" page with an AI marketing bullet and no pricing attached [29].
- Patreon and Calendly show no AI pricing structure at all, suggesting AI isn't yet monetized as a distinct line item on those platforms [22] [83].

Raw Inference Pricing: A Separate, Fast-Moving Layer

Beneath application-level pricing sits the underlying model API market, where prices are compressing quickly:

- Perplexity added NVIDIA's Nemotron 3.5 Lightning to its Agent API at $0.0115 per 1M input tokens and $0.17 per 1M output tokens, explicitly positioned as a cheap execution layer to pair with pricier frontier planning models [5].
- Google's Gemini 3.7 Flash launched at $0.75/1M input and $3.75/1M output tokens, half the cost of the prior 3.6 version, while improving coding accuracy [72].
- AWS Bedrock prices foundation models by modality, provider, and tier (on-demand, batch, Standard/Flex/Priority/Reserved), all usage-metered rather than flat-rate [42 120].
- Databricks prices its AI products by DBU consumption, starting at $0.07/DBU for both its core AI platform (AI Gateway, Agent Bricks, AI Functions) and its Genie assistant [130].
- OpenAI's Sam Altman endorsed a claim that GPT-5.4 now costs roughly one-thirteenth what its "Luna" flagship did four months earlier at matching benchmark performance, framing this cost compression as outpacing Moore's Law [28]. OpenAI has also cut prices on GPT-4 Turbo and GPT-3.5 Turbo directly [55].
- xAI's Elon Musk has repeatedly pushed comparative pricing claims: Grok 4.6 matching Claude Opus 5 Max on agentic benchmarks at $0.84 per completed task versus Claude's higher cost [109], beating GPT-5.6 Sol on a legal-research benchmark at roughly one-thirteenth the cost [138], and matching GPT-5.6 Sol's benchmark score at 5x cheaper output pricing and 2.5x cheaper input pricing [139]. These are Musk's own framings of third-party benchmark data, not independently verified company claims.

Seat-Based Pricing Is Getting More Complicated by AI Add-Ons

Traditional per-seat SaaS pricing hasn't disappeared, but AI is stacking new cost dimensions on top of it:

- Intercom's pricing is a hybrid of per-seat tiers (starting $29/seat/month) plus a separate usage-based "per Fin outcome" charge, making true cost highly variable depending on AI resolution volume, with additional a-la-carte add-ons (Pro, Copilot, Proactive Support Plus) layering further cost dimensions [2].
- OpenAI's ChatGPT Business added a Premium seat tier at $125/month (versus $25 Standard) offering 5x usage and no five-hour cap, letting teams mix seat types in one workspace [108].
- HubSpot is explicitly moving from pure subscription pricing to a hybrid of consumption-based and seat-based models, acknowledging this shift carries revenue unpredictability risk and requires new metering infrastructure, all while facing pressure from AI-native competitors and generic LLMs it says customers might use instead [24][49] .

Regional and Segment-Based Price Discrimination

- Cursor launched a $7/month tier exclusive to India, positioned below its standard $20 Pro price, to capture a price-sensitive segment through localized rather than global pricing [17][88] .

Hardware Moving to Subscription/Leasing Models

Apple's Apple Upgrade program shows the usage/subscription logic spreading beyond software into hardware: a leasing program (via Klarna) across iPhone, iPad, Mac, and Apple Watch offering lower monthly payments than purchase financing, explicitly framed internally as a shift toward recurring hardware revenue and reduced upfront friction [34][85][100][126][134][140] . Separately, Apple CEO Tim Cook has floated (not confirmed) a paid tier for advanced Siri AI features delivered through existing iCloud+ subscriptions, which would mirror the "premium compute" upsell pattern seen at GitHub Copilot and Microsoft 365 [67].

Enterprise Consumption Governance Is Becoming Standard

As credit-based pricing spreads, companies are simultaneously building governance tooling around it, suggesting the market has moved past "will we meter AI" into "how do we let customers control what they spend":

- Figma added admin controls for per-user AI credit limits with request-to-increase workflows, shipped as part of Config 2026 [62][79][110] .
- Netlify extended team-wide AI credit limits to per-member overrides, giving Team Owners two-tier spend management [4].
- Salesforce released a Data 360 usage-tracking data model refreshed every five minutes to separate metered from unmetered Agentforce consumption and estimate Flex Credit costs, closing what it calls a documented observability gap in production deployments [51].
- GitHub added per-model token breakdowns to usage reports so customers can attribute AI credit consumption by model [95].
- Anthropic's Analytics API gives Enterprise customers programmatic access to Claude usage and engagement data, aggregated per organization per day [75].

Billing Infrastructure Has Become Its Own Market

The complexity of usage-based AI pricing has spawned a dedicated infrastructure layer, rather than companies building metering themselves:

- Metronome, described as "a key billing backbone for the AI industry," powers monetization for OpenAI, Anthropic, Databricks, and NVIDIA, and was acquired by Stripe (deal closed January 2026) to become Stripe's core usage-based billing layer [19]. It has since shipped daily recurring credits and commits [26], configurable prepaid balance commit durations [74], PLG self-serve invoicing with expanded marketplace support [39], and mid-contract billing provider switching [105].

Stripe now surfaces Metronome directly in its merchant onboarding flow for new users needing usage-based billing [57].

- Orb, a competing usage-based billing platform serving SaaS and AI companies, was acquired by Adyen for $335 million in a deal that closed July 2026, continuing to operate as a standalone product [118].
- Mixpanel chose to build its billing system in-house rather than adopt an existing platform, citing the need to align pricing logic directly with its own usage metrics despite the engineering burden [90].

This pattern (three separate signals of billing-infrastructure M&A and adoption in a matter of months) suggests usage-based AI pricing has become complex enough that metering, rating, and invoicing are now treated as their own technical problem rather than a pricing-page afterthought.

Cost Optimization Is Becoming a Product Feature, Not Just a Back-Office Concern

Several sources show AI cost management shifting from "how we bill customers" to "how customers and vendors both manage token spend as a resource":

- Databricks CTO Matei Zaharia argues token consumption should be managed like traditional infrastructure resources (compute, memory, bandwidth), routing all AI through a centralized Unity AI Gateway and pushing budget controls down to individual engineers [45].
- Zapier CEO Wade Foster reported that the company's heaviest internal AI users spend up to $30,000/month on coding agents, but argued token spend is a poor proxy for productivity, citing internal research across 1,500 companies showing AI runs in only 18% of agentic workflow steps [132].
- Square published benchmarking guidance for matching model cost-tier to task type across its multi-agent product, finding a 5.5x price spread among top models but only an 8.9-point score spread among best performers, and recommending cheaper models handle high-volume execution while reserving frontier models for coordination [94].
- Perplexity assigns different models to different workload types within its Computer product (GPT 5.6 Terra as default subagent, Luna for scheduled automations) explicitly for cost-effectiveness [96].
- Harvey post-trained a smaller open-weight model (Nemotron 3.5 Lightning) on legal tasks, cutting average token output by 58% while improving performance 8.3x, and separately built "Harvey Tenet" using reward-shaped training to reduce inference-time token consumption while improving legal benchmark completion rates [78][144] .

An Emerging Model Nobody Has Fully Priced Yet: Agent-to-Business Payments

A cluster of signals points toward a pricing frontier that isn't reflected in any pricing page yet: charging AI agents themselves as transacting customers, rather than the humans who deploy them.

- Anthropic engineer Thariq Shihipar argues SaaS companies are leaving revenue on the table by not exposing their products as agent-facing APIs billed per interaction rather than per seat, calling this a structural mismatch rather than a technical limitation [131].
- Coinbase CEO Brian Armstrong argues agent payments are inevitable because autonomous agents will vastly outnumber humans and each will need payment capability, framing this as foundational infrastructure demand rather than a speculative feature [112]. Coinbase Business now accepts payments initiated by AI agents with no additional merchant setup [115].
- Stripe expanded stablecoin "machine payments" to 30 additional countries specifically for agent-initiated transactions (browser sessions, research queries, e-commerce orders) [101].
- PayPal opened MCP-based payment APIs so external AI agents can execute purchases autonomously, and released a toolkit to help merchants build commerce agents [13][35][53] .
- Vercel extended its MCP integration to support agent-initiated e-commerce purchases, not just data retrieval [86].

None of these sources show an actual per-agent-transaction price list yet; they describe infrastructure being built ahead of a pricing model that hasn't fully formed.

Notable Absences and Gaps

Several categories of company show no priced AI feature at all in these sources, which is itself informative: Calendly, Patreon, and PostHog's base pricing page list zero AI-specific pricing, and Rippling and Salesforce's top-level pricing pages disclose no concrete numbers, pointing instead to sales contact or opaque credit calculators [22][83][107][29] [10]. This suggests AI monetization maturity varies widely even among well-funded SaaS companies, and that "contact sales" remains a common way to avoid publicly committing to a usage-based number before the metering infrastructure and unit economics are fully worked out.

Sources

Launch

7 Launched AI Search, a no-code search engine for enterprise data

15 Moved user surveys to GA with usage-based pricing tied to response volume

34 Launches Apple Upgrade leasing program with lower monthly payments than purchase financing

53 Released toolkit letting merchants build AI agents to automate commerce tasks

72 Launched Gemini 3.7 Flash with 43.6% coding accuracy, half the price of 3.6

99 Launched real-time monitoring tool that auto-flags agent performance drops across channels

106 Upgraded Claude Opus 4.8 to default; adds dynamic workflows and security-guidance plugin

116 Launched Replay Vision, AI layer that auto-generates bug-fix PRs from session recordings

126 Launched Apple Upgrade hardware leasing program in the U.S.

134 Launched Apple Upgrade leasing program for iPhone, iPad, Mac, Watch via Klarna

Feature

4 Pro plans gain per-member AI credit limits with team-wide overrides

5 Made Nemotron 3.5 Lightning available in Perplexity Agent API

11 Added context-based scoring and multi-agent tracing to Agentforce Observability

13 Opened payment API to let AI agents execute purchases autonomously

20 Added AI-powered contract analysis to automate customer agreement review

26 Billing API adds daily recurring credits and commits

27 Cloud agents cut token use 20-30%, 80% on computer-use tasks

39 Added PLG invoicing and expanded cloud marketplace integrations to billing platform

51 Released AI Agent Generative AI Usage Data Model for real-time Agentforce consumption tracking

52 Previewed next-generation Apple Intelligence and redesigned Siri AI

55 Cut prices on GPT-4 Turbo and GPT-3.5 Turbo models

56 Shipped Free Mode tier and faster UI redesign

57 Stripe now recommends Metronome for usage billing in onboarding flow

62 Admins gain per-user AI credit limits and request workflows

65 Agent Runners can now start projects from natural-language prompts with built-in infrastructure

67 Cook signals paid tier for advanced Siri AI features via iCloud+

68 Retail store wait times fell 13.9% after pairing AI agents with human staff

69 Embedded subscription management into Apple Pay checkout flow

71 Prompt management exits beta with metadata and label-based fetching

74 Billing API now supports configurable prepaid balance commit duration and rollover

75 Added Analytics API for Enterprise customers to track Claude usage

79 Admins gain per-user AI credit limits and request workflows

86 Vercel MCP expanded to support e-commerce purchases

87 Metric Explorer Chart exits beta, generally available

95 Usage reports now break down token consumption by model

96 GPT 5.6 Terra becomes default subagent model in Perplexity Computer

98 Fin agent now integrates with HubSpot and Freshworks without migration

101 Expanded stablecoin machine payments to 30 additional countries for agent transactions

103 Claude Code now included in Team plan Standard seats

105 Billing platform now supports switching providers mid-contract without creating new one

108 Introduces Premium seats for ChatGPT Business at $125/month with 5x usage

110 Admins gain per-user AI credit limits and request workflows

113 Agentforce Multi-Agent Orchestration reaches general availability for routing across specialist

115 Coinbase Business now accepts payments from AI agents

117 Added Kimi K3 model option with US-based inference and zero data retention

119 Chat SDK now supports Claude Managed Agents execution

127 Square sellers can accept orders placed through Google Maps agents

128 Added Claude Sonnet 4.6 model option to Windsurf IDE

142 Claude Cowork gains OpenTelemetry monitoring support

147 Added Claude Cowork to Analytics API

149 CX Agent gains automated price-matching and refund processing

Internal tool

3 Built framework measuring AI agent outcomes beyond conversation quality

46 Multi-agent research system uses orchestrator-worker pattern with parallel subagents

70 Outlined blueprint for AI agents that autonomously purchase products on customer behalf

84 Built Unified Planner reasoning engine powering all new Agentforce agents

90 Built custom billing system to tie pricing directly to usage metrics

91 Engineering blog details MCP code execution approach for token-efficient agents

94 Benchmarked optimal model-tier assignments for multi-agent workflows in Buzz

122 Published framework for measuring AI coding-agent performance beyond standard benchmarks

141 Released framework for grounding AI agents in real-world business operations

AI adoption

12 Negotiates multiyear publisher deals worth nine figures to feed current news into Siri AI

25 Grok Bot autonomously recovered churned customers and covered its own cost

45 Token budgeting becomes core infrastructure for managing AI spend at scale

Org

125 Salesforce to acquire Fin for $3.6B

Earnings

21 Annualized revenue surges to $65B; investors project $100B–$120B finish for 2026

24 HubSpot shifting to consumption-based and seats-based pricing while facing intensifying AI

40 Q2 revenue doubles to $11.6B; outpaces OpenAI's $6.7B quarterly sales

Research

78 Post-training on legal tasks boosts smaller model performance 8.3x while cutting token output by

144 Post-trained Harvey Tenet model doubles task completion on legal benchmarks

Perspective

6 CEO argues AI's power centralization is not structural, disputes scaling-law inevitability

28 Altman endorses pricing collapse claim: GPT-5.4 now costs one-thirteenth of Luna flagship

44 Co-founder signals interest in open-weight models for agent workloads

47 Half of major software companies won't survive five years without intent-first design

64 CEO: AI tools enable building products previously impossible to ship alone

81 AI ubiquity makes blanket 'no AI' bans incoherent, CEO argues

97 Video will be the final interface for all software, CEO predicts

104 CEO backs claim that coding harness solves ARC-AGI nearly completely

109 Grok 4.6 ties Claude Opus 5 Max on agentic tasks at half the cost per task

112 Armstrong: agent payments will inevitably drive crypto adoption as AI outnumbers humans

124 CEO: SOC2 certifications inadequate for AI-driven security threats

131 Anthropic engineer argues SaaS companies leave revenue on table by not exposing APIs for

132 Top internal AI users spend $30K monthly; most skip AI in agentic workflows

133 Frontier AI will scale reasoning via automated symbolic world models, not raw scale

138 Grok 4.6 ranks #2 in legal-research tasks, costs 13× less than GPT-5.6 Sol

139 Grok 4.6 priced 5× cheaper than GPT-5.6 Sol at equal performance

145 Armstrong: adopting AI into old workflows misses the point — work itself must change