Pricing Changes to AI Features Across the Market

Hi Alex, here's your personalized Deep Research Briefing on **Pricing Changes to AI Features Across the Market**.

Quick summary: Across the market, companies are shifting AI from opaque credits to transparent usage controls, bundling features into base plans instead of selling separately, and cutting per-token costs as inference becomes cheaper.

Companies are moving AI pricing in several distinct directions at once: cutting per-token costs as inference gets cheaper, shifting from opaque credit systems to transparent quotas, adding governance tools to control runaway spend, and testing regional or tiered pricing to capture price-sensitive segments. Below is a breakdown by category.

Structural Pricing Model Changes

Several companies have rebuilt their AI pricing from the ground up rather than just adjusting numbers.

Company	Change	Details
Zed	Prompts to tokens	Cut Pro tier 50% to $10/month with $5 included token credits, replacing flat prompt limits with usage-based billing
Hex	Seats to credits	Moved Hex Agents from flat-rate seat licensing to monthly credit grants (10 to 60 credits by tier) with pooled add-ons and auto top-ups
Codeium (Windsurf)	Credits to quota tiers	Replaced an opaque credit system with transparent Free, Pro, Teams, and a new Max tier for power users
Netlify	Per-seat to flat-rate	Eliminated per-seat charges on Pro plan in favor of $20/month flat rate with unlimited seats, while recalibrating consumption meters on bandwidth, compute, and requests to keep 98% of customers' bills flat or lower
Neon	Removed billing floor	Eliminated the $5 monthly minimum, moving to pure consumption-based pricing with no baseline commitment

Linear	Coding sessions repriced	Shifted from bundled pricing to passing through provider token rates with no markup, plus a flat $0.25 per 20 minutes of runtime and optional spend caps

Price Cuts on Underlying Models

Frontier labs have been cutting per-token prices sharply, often while holding or improving capability.

- Anthropic cut cache read pricing 75% to $0.25 per 1M tokens while keeping Fable 5.1 input/output rates unchanged at $10/$50 per 1M tokens.
- Anthropic's Claude Opus 5 matches Fable 5 performance at roughly half the price and became the new default for Claude Max subscribers.
- Google cut Gemini 3.7 Flash pricing 50% to drive API adoption and developer wallet share. Separately, Gemini 3.6 Flash and 3.5 Flash-Lite launched with lower per-token pricing ($1.50/$7.50 per 1M tokens and $0.30/$2.50 per 1M tokens respectively) alongside efficiency gains.
- Elon Musk endorsed a claim that Grok 4.6 undercuts GPT-5.6 Sol by 5x on output tokens ($6 vs $30 per 1M) and 2.5x on input tokens ($2 vs $5 per 1M) at equal benchmark performance.
- Greg Brockman backed benchmark results showing GPT-5.6 Luna held its ARC-AGI performance after an 80% price cut, with cost per task dropping to $0.18 (ARC-AGI-2) and $0.07 (ARC-AGI-1).
- Vercel's Production Index data shows average price per token on its AI Gateway fell 13.6% in July and 23.2% in August, driven by usage shifting toward cheaper open-weight models like DeepSeek V4 Flash rather than direct price reductions from labs.
- 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, aimed at high-volume execution tasks.

Free Tiers and Bundling AI Into Existing Plans

A number of companies have moved AI features into base pricing rather than charging separately.

Company	Change

Anthropic	Added Claude Code to all Team plan Standard seats at no additional charge
Granola	Bundles AI meeting notes and AI chat into the free Basic tier, with higher tiers unlocking more advanced models rather than charging separately for AI
Klaviyo	Includes an AI-powered subject line generator and 10,000 Composer credits in the free plan
Greenhouse	Bundles AI Notetaker and Talent Matching into its base Core tier
Smartsheet	Includes Smart Assist and AI-powered dashboards starting at the lowest paid tier ($9/user/month)
Shopify	Bundles AI website builder, tokens, and Sidekick assistant into every tier including entry-level Basic, plus a $0-fee Agentic tier for AI-channel selling
Webflow	Includes Webflow AI in both the free Starter and Basic tiers
Lattice	Bundles an AI Agent into all base products at no extra cost
Retool	Includes 250 AI credits/month in the free tier, scaling to 10,000 on Enterprise

Credit and Usage Governance Tools

As AI features shift to metered consumption, companies are adding controls to prevent bill shock and manage spend.

- Figma repeatedly rolled out admin controls letting teams set custom per-user AI credit limits with request workflows for increases, reported across multiple updates through mid-2026.
- Figma separately increased AI add-on credits included with subscriptions: 2x for Pro plan, 1.6x for Org and Enterprise, applied immediately to new and existing purchases.
- Anthropic launched Smart Reports in beta for Claude Enterprise, analyzing usage, costs, and workflow friction points, and earlier added an Analytics API for Enterprise customers to track usage programmatically.
- Salesforce released an AI Agent Generative AI Usage Data Model, refreshed every five minutes, letting admins separate metered from unmetered Agentforce usage and estimate Flex Credit costs.
- GitHub added per-model token visibility to usage reports, breaking down input, output, and cache consumption by model, and gave Copilot users three tiers

(efficiency, balance, intelligence) to tune cost-quality tradeoffs in automatic model selection.

- Netlify's Agent Runners now scope ambitious prompts to fit available credit budgets rather than cancelling runs mid-execution when credits run out.
- Metronome moved token billing to public preview, automating LLM pricing from 30+ inference providers via Stripe's model API so companies can apply markup without manually updating rate cards as provider prices shift.

Capacity and Rate Limit Changes

- Anthropic is permanently raising Claude Code weekly limits 25% for Pro, Max, Team, and seat-based Enterprise plans starting September 14, though this represents a net 17% reduction from a temporary 50% boost that is expiring.
- OpenAI restored 5-hour daily limits on Codex and Work for ChatGPT Plus after operating without restriction for several weeks, citing compute load management.
- Meta raised Muse Code's contributor tier usage limits substantially and reset accumulated usage for all active subscribers.
- Neon raised included monthly data transfer from 100 GB to 500 GB across all paid plans, applied automatically to existing customers.
- Airtable increased the work performable per AI credit 5x without changing credit pricing, passing through underlying model cost and capability improvements.

Regional and Segment-Specific Pricing

- Cursor launched Cursor Start at roughly $7/month exclusively in India, its third-largest market, sitting below the standard $20/month Pro tier as a localized play for a price-sensitive region.
- OpenAI added a $125/month Premium seat tier within ChatGPT Business offering 5x usage and no hourly limits, allowing mixed-tier workspaces to capture higher-spending users within the same product.
- OpenAI is giving free GPT-6 Astra Pro access to verified U.S. clinicians through ChatGPT for Clinicians [1], and providing free frontier model access (including GPT-5.6) to 100,000 academic researchers through 2027.

- SpaceXAI expanded Grok Bot access to SuperGrok Plus, Cursor Pro+, and Cursor Teams subscribers with a free trial for others, and separately offered Grok Bot for Enterprise free for two weeks to Grok and Cursor enterprise customers.
- Apple signaled a shift toward monetizing advanced Siri AI capabilities through iCloud+ subscription tiers rather than bundling them universally.

New Tiers for Teams and Enterprises

- Anthropic opened self-serve Enterprise plans without requiring sales engagement, bundling Claude, Claude Code, and Cowork into a single seat type.
- Replit launched a $100/month Pro tier for serious builders with tiered credits, credit rollover, and up to 15 team members, while repricing the Core plan lower with new cost controls. CEO Amjad Masad later framed Replit's separate Free Mode as an accessibility move, saying he found it "painful" when users were priced out of AI coding.
- Runway launched a Team Plan at $69/seat/month ($55 annually) for 2 to 9 person teams with pooled credits (6,900 per seat) and shared projects, and separately unified its credit system so enterprise customers can use one credit pool across both its web app and Dev API.
- Zed introduced its first tiered pricing with a $30/seat/month Business tier adding admin controls and monetizing AI features that were previously unmonetized under a cost-plus model.
- Vercel launched Flat Rate CDN for Pro teams as an alternative to usage-based billing, with a default $20/month tier covering 1M requests and 1TB transfer, scaling to $300/month for 150M requests and 50TB, explicitly designed to prevent viral-traffic bill shock.
- Tableau introduced capacity-based pricing for Tableau Cloud (purchasing blocks of Viewer access shared across employees, partners, and AI agents rather than per-seat) alongside new Cloud+/Server+ editions bundling agentic analytics.
- Netlify added higher Pro plan tiers (10,000, 15,000, 20,000 monthly credits) with lower effective per-credit costs and rollover on the three higher tiers, aimed at capturing more predictable revenue from power users.

Paid Add-Ons Versus Bundled AI

Pricing strategy varies by company on whether AI is a separate line item or folded into the base plan.

Company	AI pricing approach
Sentry	Seer AI Debugger is a $40 per active contributor/month add-on to Team, Business, or Enterprise plans [2]
GitLab	Duo Agent Platform bundles limited credits into Premium ($12/user/month) and Ultimate ($24/user/month), then bills additional usage at $1 per GitLab Credit [3]
Notion	AI chat and Meeting Notes are limited trials on Free/Plus and fully included on Business/Enterprise, but Custom Agents and Workers require paid Notion credits at $10 per 1,000 monthly credits [4, 5]
Ahrefs	Tracked AI prompts and Brand Radar are billed via metered overage rates ($0.010 to $0.020 per check), with Brand Radar AI starting at $199/month [6]
Atlassian	Every tier includes a fixed Rovo credit allotment (25 Free, 70 Standard, 150 Premium) scaling by tier rather than unlimited inclusion [7]
Airtable	Every tier includes base AI credits (500 to 25,000 per user/month), with additional packs starting at $20/month for 10k credits [8]

Vendor Economics Behind the Price Moves

Some of the underlying cost dynamics driving these pricing changes:

- OpenAI absorbed a 20% compute overhead from new monitoring and security safeguards rather than passing the cost to customers.
- Microsoft replaced OpenAI's image-generation models with its own proprietary MAI models in PowerPoint and Bing, with Mustafa Suleyman citing an 85% cost reduction for running the models in PowerPoint.
- Microsoft's GitHub Copilot introduced HydraFusion, which coordinates multiple models across planning, building, critique, and completion phases, cutting cost-to-outcome by up to 67% versus single-model baselines, according to comments attributed to CEO Satya Nadella.
- Vercel's CEO Guillermo Rauch argued that OpenAI's Sol price cuts made it the fastest-growing frontier model on Vercel's AI Gateway, contending that demand for AI inference is highly price-elastic and that routing gateways are becoming essential infrastructure for capturing cost volatility.

- Netlify's CEO Mathias Biilmann argued that the price of AI intelligence shapes product viability as much as raw capability, framing pricing strategy as a primary competitive factor rather than secondary to capability.
- Salesforce launched Koa, a reasoning model co-developed with Nvidia by post-training the open-weight Nemotron base model on synthetic data, positioned as a token-efficient, data-sovereign alternative to frontier models for sales and support tasks within Agentforce.

Together these moves point to a market where AI features are increasingly billed by consumption rather than flat fees, where labs are competing aggressively on price at equal or near-equal capability, and where platforms are racing to add governance tools so that usage-based pricing does not turn into unpredictable bills for their customers.

Read the full report
