

THE
SPECTATOR
YOUR LENS INTO THE WORLD OF AI

The Open-Prem Inflection Point V4

October 2026 Update

The evolution of self-hosted enterprise infrastructure since spring, alongside the strategic impact of agent-generated software architectures.

By: David Borish, AI Strategist

www.TheAISpectator.com | www.DavidBorish.com

Table of Contents

Executive Summary.....	3
The Model Landscape: Trillion-Parameter Open Weights.....	4
The open-weight leaders, October 2026.....	4
What changed by family.....	5
The licensing shift.....	6
Usage: open-weight models now carry most OpenRouter traffic.....	6
AI-Built Software and the Application Layer.....	7
What happened in October.....	7
Enterprises were already moving.....	8
What I expect next.....	8
What will not change quickly.....	9
A build, adopt, or buy guide.....	9
Agents: The Skill Registry Grew Faster Than Review.....	10
OpenClaw.....	10
ClawHub: growth outpaced review.....	10
NemoClaw.....	11
Remaining gaps.....	11
Hardware Economics: Prices Rose.....	11
What this means.....	12
Updated deployment tiers.....	12
The Compliance Imperative.....	13
EU AI Act: high-risk deadline moved to December 2027.....	13
IBM Cost of a Data Breach 2026.....	13
GDPR.....	14
Total Cost of Ownership: What You Compare Against Decides the Answer.....	14
Current API prices (per million tokens).....	14
A worked payback example.....	15
Updated decision framework.....	15
The Open-Source Creative Stack.....	16
Conclusion.....	17
About the Author.....	17
Sources.....	18

Executive Summary

Six months ago, V3 argued that open-source models, autonomous agents, and on-premises hardware had become a complete enterprise AI architecture. That argument still holds. Two things have changed enough to require a new version: the open-weight frontier moved again, and AI coding reached the point where a single developer could build alpha-stage reimplementations of Photoshop and Word within weeks.

The models. The best open-weight model on the Artificial Analysis Intelligence Index is now Xiaomi's MiMo-V2.6-Pro, a 1.02-trillion-parameter model released under MIT on September 22. It scores 46, against 58 for Claude Opus 5.5, the top proprietary model. GLM-5.3, Kimi K3, and StepFun's Step 5 follow within two points. Kimi K3 is 2.8 trillion parameters. Qwen3.8 is 2.4 trillion. Meta released two open-weight models: the 30B Muse Glimmer under Apache 2.0 in August, and the 816B Muse Spark Open under a custom license on October 15. Muse Spark Open, at 39, is now the highest-scoring US open-weight model. NVIDIA shipped Nemotron 3 Ultra and, on October 15, Nemotron 3.5 Super VL.

The usage. On OpenRouter, the largest public model-routing marketplace, open-weight models passed closed models in token volume in early June and held about 60% of tokens by mid-June. I found no earlier sustained majority. Closed labs still collect most of the money: on Vercel's AI Gateway, open-weight models ran 29% of June tokens on under 4% of spend.

One trend runs against the V3 thesis. Several of the strongest open models now ship under custom licenses with revenue thresholds, security reviews, or branding requirements. "Open weights" and "permissive license" are no longer the same thing, and enterprises have to read each license.

The build era. In early October, developer Brandon Thomas used Claude Opus 5.5 to build open-source reimplementations of seven Adobe applications and then Word, Excel, and PowerPoint, all in Rust, all under MIT or Apache 2.0. He describes them as clean-room. The suite is called ArtCraft, and its Photoshop equivalent, PhotoCraft, passed 64,700 GitHub stars within weeks. The apps are early alpha and not ready for professional work. Separately, McKinsey's 2026 State of AI survey found that 32% of organizations had decided against buying at least one software product or feature because they could build it in-house with agentic coding tools.

This paper adds a new argument. As coding models improve, I expect companies to build more of their own software. They would start from a growing base of AI-built, permissively licensed open-source applications, modify them for their own workflows, and keep the result private. The Open-Prem thesis has always been about owning the model and the data. This edition extends it to the application layer.

The economics. Hardware prices rose. H100 rental was up about 20% for the year by May, and standard RTX 5090 cards sold for \$4,930 to \$5,090 in early September against a \$1,999 list price. Hosted open-weight APIs cost far less than proprietary ones, though prices vary widely: MiMo-V2.6-Pro costs \$0.435 per million input tokens and \$0.87 per million output tokens, Kimi K3 costs \$3.00 and \$15.00, and GPT-6 Astra costs \$10 and \$50. Self-hosting now wins on control and compliance more often than on price, except at high volume.

The compliance clock. The EU adopted the Digital Omnibus in July, moving standalone high-risk obligations to December 2, 2027. Transparency obligations under Article 50 took effect on August 2,

2026 as planned. IBM's 2026 breach report put the global average breach at \$4.99 million and found shadow AI involved in 43% of breaches, up from 20%.

The Model Landscape: Trillion-Parameter Open Weights

In April, the largest open model I covered was DeepSeek's expected one-trillion-parameter V4. In October, four released open-weight models exceed one trillion parameters, a fifth (Mistral Large 4) is due this month, and the top four open-weight models on the Artificial Analysis leaderboard are all from Chinese labs. The highest-scoring US open-weight model is Meta's Muse Spark Open at 39, released October 15 under a custom license. Reflection's Beam (30) and NVIDIA's Nemotron 3.5 Super VL (26) follow among US models, per [Artificial Analysis](#).

A note on scores: Artificial Analysis has rescaled its Intelligence Index twice since V3. Kimi K3 scored 57 on v4.1 and 44 on v4.3, and Meta's Muse Glimmer fell from 35 at its August launch to 17 today. Scores below are from the Artificial Analysis leaderboard on October 16, 2026 (Index v4.3.2). They should not be compared with numbers in earlier editions of this paper or in launch-day articles.

The open-weight leaders, October 2026

Model	Lab	AA Index (Oct 16)	Total / active params	Context	License
MiMo-V2.6-Pr o	Xiaomi	46	1.02T / 42B	1M	MIT
GLM-5.3	Z.ai	45	744B / 40B	1M	Custom (security review above \$10B revenue)
Kimi K3	Moonshot AI	44	2.8T / 104B	1M	Custom (branding above 100M MAU or \$20M monthly revenue)
Step 5 Preview	StepFun	44	600B / 27B	1M	Custom StepFun Model License (separate agreement for device makers and for inference providers above \$20M revenue)
GLM-5.3-Flash	Z.ai	42	320B / 18B	1M	MIT
Qwen3.8-2.4T-A95B	Alibaba	40	2.4T / 95B	262K, extendable to 1M	Custom Qwen3.8-Max license

Model	Lab	AA Index (Oct 16)	Total / active params	Context	License
DeepSeek V4.1-Flash	DeepSeek	39	552B backbone / 16B	1M	MIT
Muse Spark Open	Meta	39	816B / 42B	1M	Custom Muse Model License (agreement for model-as-a-service above \$20M revenue; "Built with" notice required)
MiMo-V2.6-Flash	Xiaomi	38	309B / 15B	1M	MIT
DeepSeek V4 Pro	DeepSeek	36	1.6T	1M	MIT
Nemotron 3.5 Super VL	NVIDIA	26	120B / 12B	1M	NVIDIA Open Model
Muse Glimmer	Meta	17	30B dense	128K	Apache 2.0
Granite 4.2 30B	IBM	13	30B dense	131K	Apache 2.0

Source: [Artificial Analysis leaderboard](#), October 16, 2026, and model cards. Mistral Large 4 (1.05T total, 49B active) entered public preview on October 6 and scores 38; its weights are promised for the end of October.

What changed by family

DeepSeek. [V4 shipped on April 24](#) at 1.6 trillion parameters under MIT, larger than the pre-release reporting I cited in V3. On September 10 DeepSeek released [V4.1-Flash](#) under MIT. It scores above V4-Pro on the Artificial Analysis index, and DeepSeek now routes V4-Pro API requests to it.

Alibaba Qwen. Qwen3.8 is the first Max-class Qwen model with downloadable weights. The 2.4T checkpoint carries a custom license that [requires a separate commercial agreement](#) for large organizations offering AI services. The smaller Qwen3.8-27B stays under Apache 2.0.

MiniMax. M2.7, which I described in V3 as proprietary, was later released as open weights with [commercial-use restrictions](#). [MiniMax M3](#), released June 1, followed with open weights under the custom MiniMax Community License, a 1M context window, and native multimodality. Its API costs [\\$0.30 input and \\$1.20 output per million tokens](#) for prompts up to 512K tokens.

Meta. Llama 4 is no longer Meta's lead line. Meta Superintelligence Labs released the closed Muse Spark in April. On August 10 it released [Muse Glimmer](#), a 30B model under Apache 2.0, which carries none of the EU restrictions V3 flagged in the Llama 4 license. On October 15 it released [Muse Spark Open](#), an 816B model with 42B active parameters. Its custom Muse Model License requires companies with more than \$20 million in annual revenue to sign a separate agreement before offering the model as a service, and requires every product using it to display a "Built with Muse Spark Open" notice.

Mistral. [Mistral Large 4](#) is a 1.05-trillion-parameter model trained on 3,800 Grace Blackwell GPUs in Mistral's own European data centers. The preview scores 38. For EU enterprises focused on sovereignty, it will be the largest European open-weight option once weights ship.

Z.ai. GLM-5.2 shipped under MIT in June. GLM-5.3 moved to a custom license. GLM-5.3-Flash, a new 320B multimodal base model, stays under MIT.

NVIDIA. [Nemotron 3.5 Super VL](#) (120B, October 15) scores 26, ahead of the larger [Nemotron 3 Ultra](#) (550B, announced May 31) at 23. [Nemotron 3.5 Bolt](#), a 74B model with 9B active parameters, is sized to run in 24GB. NVIDIA continues to publish training data and recipes, which matters for regulated buyers who need to document provenance.

IBM. [Granite 4.2](#) adds reasoning to 3B, 8B, and 30B dense models under Apache 2.0. The 30B scores 57.0 on SWE-bench Verified in IBM's testing.

The licensing shift

In V3 I wrote that most frontier open models used MIT or Apache 2.0. That is now true of MiMo-V2.6, GLM-5.3-Flash, DeepSeek V4 and V4.1, Muse Glimmer, Granite, and the smaller Qwen models. It is not true of GLM-5.3, Qwen3.8-Max, Kimi K3, Step 5, MiniMax M3, or Muse Spark Open. Each of those allows self-hosting but attaches conditions that bind large companies. Legal review of the model license is now a required step in any Open-Prem deployment, as it has long been for open-source software.

Usage: open-weight models now carry most OpenRouter traffic

Since V3, open-weight models have passed closed models in token volume on OpenRouter, the largest public model-routing marketplace. In mid-March 2026, closed labs took about 59% of OpenRouter tokens and open-weight labs 41%. On a seven-day average, open-weight models moved ahead on June 4, and by June 17 the split was 60% to 40%, according to [Dirac's compilation of OpenRouter daily usage](#). OpenRouter's own study had put open-weight usage at about a third of the platform in late 2025. I found no earlier period in which open-weight models held a sustained majority.

The pattern held into October. In OpenRouter's [rankings for the week ending October 15](#), DeepSeek V4.1-Flash led with 37.1 trillion tokens. Seven of the ten most-used models are open-weight, two (GPT-6 Luna and Claude Opus 5.5) are closed, and I could not confirm published weights for the tenth, Tencent's Hy4 preview. A second gateway shows the same direction: on [Vercel's AI Gateway](#), open-weight models ran 29% of tokens in June, up from 11% in April.

Gateway	Open-weight share of tokens	Open-weight share of spend	Period
OpenRouter	41% to 60%	Not published	Mid-March to mid-June 2026
Vercel AI Gateway	11% to 29%	Under 4%	April to June 2026

These figures measure hosted API traffic, and four limits apply. Token share is a different measure from spend: on Vercel, the top four US labs took 95% of June spend. OpenRouter is a price-sensitive marketplace, and the large closed labs sell most of their volume through their own APIs and consumer apps, which it does not see. Free variants add volume; Nemotron 3 Ultra's free tier is in the top ten. And the open-weight side is almost entirely Chinese labs, led by DeepSeek, Z.ai, Xiaomi, and MiniMax. Anthropic held about 20% of OpenRouter tokens through the crossover, which fits the hybrid pattern in this paper: open models for volume, a frontier model for the hardest tasks.

For Open-Prem, the result shows that developers now trust open-weight models with production traffic, and that most of this traffic runs on weights an enterprise could host itself. It does not show enterprises self-hosting, since this usage runs on cloud APIs. For regulated buyers who cannot send data to the model makers' own APIs, especially Chinese-hosted ones, self-hosting the same weights is how they get access to these models.

AI-Built Software and the Application Layer

AI coding models can now produce alpha-stage reimplementations of complex commercial software in weeks. As they improve, I expect enterprises to build more of their own applications on top of permissively licensed open-source code, then keep what they build private. This is the same logic as Open-Prem applied to the application layer.

What happened in October

In early October, developer Brandon Thomas announced on Hacker News that "software is over." He had used [Claude Opus 5.5 to build reimplementations of Adobe's applications in Rust](#), which he describes as clean-room. He said his motivation was an Adobe cancellation fee. In March, Adobe [settled a Justice Department lawsuit over hidden termination fees](#) by paying \$75 million and providing \$75 million in free services, while denying wrongdoing.

Within about a week he added [Word, Excel, and PowerPoint equivalents](#). The apps are released together under the ArtCraft name, free and open-source, under MIT or Apache 2.0. Thomas describes himself as a 15-year software industry veteran with roughly a decade of Rust experience.

App	Reimplements	Reported status (mid-October)
PhotoCraft	Photoshop	Early alpha. 64,700 GitHub stars, 11,300 forks, 2,429 commits, 1,700+ tests. Re-saves 307 of 309 test PSD files with matching render.

App	Reimplements	Reported status (mid-October)
WordCraft	Word	Alpha. 88% of ribbon commands implemented; reads and writes .docx.
GridCraft	Excel	Alpha. Developer reports 89% of ribbon commands and about 98% of worksheet functions.
DeckCraft	PowerPoint	Alpha. Developer reports 79% parity across 191 features.
VectorCraft, FilmCraft, LightCraft, EffectCraft, DesignCraft, PrintCraft	Illustrator, Premiere Pro, Lightroom, After Effects, InDesign, Acrobat	Early alpha; users report missing features and bugs.

The PhotoCraft README is direct about its limits: it is "not yet a Photoshop replacement for daily professional work," with gaps in generative features, 15 missing tools, typography, and plug-in support. Reviews have been mixed.

Two design choices matter more than the parity numbers. First, every PhotoCraft action is one of more than 500 commands that can be called from the UI, a CLI, a JSON control channel, or an MCP server. The software was built to be operated by agents as much as by people. Second, it runs natively on macOS, Windows, Linux, FreeBSD, and in the browser, with no account and no cloud dependency.

Enterprises were already moving

ArtCraft is the most visible example. Survey data shows the shift started earlier:

- McKinsey's State of AI 2026 survey of 1,719 respondents, published August 25, found that 32% of organizations had decided against buying at least one software product or feature because they could build it in-house with agentic coding tools. In technology companies the figure was 41%. McKinsey's report is summarized [here](#).
- Retool's [2026 Build vs. Buy report](#) surveyed 817 builders, mostly Retool customers. 35% had replaced at least one SaaS tool with a custom build, and 78% expected to build more of their own tools in 2026. The sample is people who already build software, so it likely overstates the rate for companies in general.
- GoDaddy told investors it is [testing internally built replacements for smaller third-party SaaS tools](#). West Monroe's chief people officer told [Business Insider](#) that buying the payroll-checking and workforce-insight tools its employees built with ChatGPT and Codex would have cost more than \$300,000 a year.

The coding capability needed to do this no longer sits only in closed models. On DeepSWE v1.1, [Xiaomi's published tables](#) show DeepSeek V4.1-Flash at 74.2 and MiMo-V2.6-Pro at 71.9, against 74.0 for both Claude Opus 5 and GPT-6 Astra. These are vendor-reported figures, and Xiaomi compared against Opus 5 rather than the current Opus 5.5. A company can run the build itself on self-hosted models and keep its source code and specifications inside its own network.

What I expect next

I expect AI-built, permissively licensed reimplementations of common commercial software to keep appearing. Thomas has said he expects other developers to follow, and the cost of trying is now low. Enterprises are unlikely to give an alpha Photoshop equivalent to their design teams. More likely, they will take components that fit, such as a document engine, a spreadsheet calculation core, or a PDF pipeline, and build internal tools on them. Coding agents can then adapt those tools to the company's own workflows, data models, and approval rules, and the company can keep the modified software private.

The base's license decides whether the company can keep the result private. MIT and Apache 2.0, which ArtCraft uses, permit closed derivatives. The AGPL, used by OnlyOffice, requires releasing source code for modified versions offered to users over a network.

There is a second reason to keep this code closed. Plagiarism Today argues that [code generated by AI without a human author is likely not protected by copyright at all](#). If that view holds, a company cannot rely on copyright to protect AI-written internal software, and its practical protection is to keep the code unpublished and treat it as a trade secret.

This fits the rest of the Open-Prem architecture. Agent-ready software that exposes every function as a command can be operated by the OpenClaw-style agents described later in this paper, running on local hardware, against data that never leaves the building.

What will not change quickly

- **The last 10%.** Getting an alpha to professional reliability is most of the work. Plagiarism Today cites the [ninety-ninety rule](#): the remaining 10% of the code takes the other 90% of the time. Thomas's own estimate moved from parity "within a month" to months.
- **Maintenance.** Every internal tool needs an owner. One analysis of the Retool data estimates [12.5 to 25 hours of weekly maintenance for a team replacing five SaaS tools](#).
- **Governance.** In Retool's survey, [60% of builders had built something outside IT oversight in the past year](#), and 49% had not yet shipped working software with AI. Ungoverned internal software is the application-layer version of shadow AI, which IBM now links to 43% of breaches.
- **Legal exposure.** Copyright protects code, not functionality, so [Plagiarism Today](#) judges the apps to be on decent footing against copyright claims. It also notes that Adobe and Microsoft terms of service bar studying their software's inputs and outputs in order to recreate it, that Adobe holds thousands of patents, and that one-to-one interface clones could draw trademark confusion claims. It does not expect Adobe to act quickly.
- **Mature alternatives exist.** LibreOffice and OnlyOffice remain the stable choices for office work today. GIMP, Krita, and Affinity have existed for years without displacing Adobe.

A build, adopt, or buy guide

Software type	Recommended approach	Reason
Internal admin tools, dashboards, workflow apps	Build with coding agents	Low complexity, high fit, the category surveys show moving first
Workflows that differentiate the business	Build on a permissive open-source base, keep closed	Encodes how the company works; trade secret is the protection
Commodity desktop apps (office, image, PDF)	Adopt mature open-source now; pilot AI-built suites	LibreOffice and OnlyOffice are stable; ArtCraft-style apps are alpha
Systems of record (ERP, payroll, core banking, EHR)	Keep buying	Compliance, audit, and integration depth outweigh license savings
Anything built with AGPL code	Plan to publish, or choose another base	Network use triggers source release

Agents: The Skill Registry Grew Faster Than Review

OpenClaw kept growing, its skill registry roughly quadrupled, and the security problem I described in V3 did not go away. NemoClaw is still pre-1.0. The agent layer is usable for production only with an internal, vetted skill catalog.

OpenClaw

Two OpenClaw releases matter for enterprise use. [Version 2026.3.28](#) added plugin approval hooks, which let a plugin pause a tool call and wait for human confirmation. [Version 2026.4.7](#) added a persistent memory-wiki and session branching and recovery. Creator Peter Steinberger [joined OpenAI in February](#); the project moved to a foundation and continues as open source.

The hybrid pattern from V3 still applies: a cloud model supervises, local open-weight models execute. What changed is the local tier. With MiMo-V2.6-Flash, GLM-5.3-Flash, and DeepSeek V4.1-Flash available as open weights, the supervisor can also run locally for organizations that need full sovereignty, at the cost of data-center-class hardware.

ClawHub: growth outpaced review

Measure	Earlier figure	Latest figure
Skills in registry	13,700+ (cited in V3)	About 52,700 (June)
Malicious skills found	Koi Security: 341 of 2,857 sampled (February)	RankClaw: 1,103 of 14,706, full registry (March)
Skills flagged by scanners	Snyk: 13.4% of 3,984 with a critical issue (February)	gradually.ai: 48.4% of 16,797 with at least one finding of any severity
Mandatory security review	No	Still no; each new submission is scanned with VirusTotal

The audits used different methods and samples, so the figures are not directly comparable with each other.

Tencent's security team scanned about 50,000 skills in April and found [samples that passed ClawHub's checks](#). OWASP now maintains an [Agentic Skills Top 10](#), which lists malicious skills as the first risk. My V3 advice stands: read ClawHub skills as source material and do not install from the registry directly. Run only skills that are bundled, written internally, or rebuilt internally from reviewed source, and keep them in a private registry.

This connects to the application-layer argument. A company that has coding agents write its own skills from a public reference, rather than installing third-party code, is applying the same approach at the agent layer.

NemoClaw

NemoClaw remains in early preview, now at [version 0.0.117 and later](#). OpenShell is now the single source of sandbox network policy. NVIDIA added managed MCP, DGX Station and DGX Spark setup, and [support for agent frameworks beyond OpenClaw](#), including Hermes and LangChain Deep Agents. That last change matters: NemoClaw now works as a general sandbox layer for several agent frameworks. Nemotron 3.5 Super VL, released October 15, [runs on a single DGX Spark in NVFP4](#), so a fully local NemoClaw deployment with a multimodal model fits on a desktop-class system.

Remaining gaps

The gaps I listed in V3 (compliance monitoring, ERP integration, HR, ticketing, SOX-grade financial controls) are still open. They are also the categories where the build-era approach is most useful: these are internal workflows with company-specific rules, a good fit for software a company writes and keeps.

Hardware Economics: Prices Rose

In V3 I wrote that older GPUs were getting cheaper and that organizations did not need the latest hardware to start. That trend reversed. Demand for AI compute pushed prices up across generations, and the main new option for small on-prem deployments is Apple's M5 Ultra.

Hardware	V3 (April)	October 2026
H100 cloud rental	Under \$3/hr and falling	Up about 20% for the year as of May; A100 up nearly 15%
Blackwell B300	Shipping since January	DGX B300 lead times of 8 to 12 weeks
NVIDIA Vera Rubin	Not covered	Full production June 1; first NVL72 rack running at CoreWeave June 1; 288GB HBM4 and 50 PFLOPS NVFP4 per GPU in NVIDIA's published specs; volume ramp in Q4 2026, with shipments growing through the first half of 2027
AMD Instinct	MI350 shipping; MI400 "2026"	MI450 sampling, production in H2 2026 : 432GB HBM4, 19.6 TB/s; MI500 in 2027
RTX 5090	\$2,900 to \$5,000+ street	\$4,930 to \$5,090 for standard cards in early September, then sold out at first-party US retailers by mid-September ; MSRP still \$1,999
Mac Studio	M3 Ultra, up to 512GB	M5 Ultra, up to 512GB at 1.2 TB/s, from \$5,499; 512GB configuration due late October, price not announced

What this means

For data-center deployments, Rubin is shipping, but allocation goes to hyperscalers and frontier labs first. An enterprise that needs capacity before mid-2027 should plan on B300 or MI450. The useful development for Open-Prem is memory: Rubin at 288GB and MI450 at 432GB per GPU mean trillion-parameter open models fit on fewer accelerators.

For small deployments and agent fleets, the RTX 5090 is poor value for AI work at street prices about 2.5 times its list price. The M5 Ultra Mac Studio offers up to 512GB of unified memory at 1.2 TB/s in one desktop unit. Apple claims up to 4.3 times the peak AI compute of the M3 Ultra, and multiple units can be clustered over Thunderbolt 5 with RDMA into one memory pool. Independent benchmarks were limited at the time of writing, so test on your own models before buying.

For model selection, the leading open models are now too large for a single workstation. Kimi K3 needs about 1.4 TB to run. Smaller releases fit on one workstation or GPU: Qwen3.8-27B, Muse Glimmer 30B, Granite 4.2 30B, and Nemotron 3.5 Bolt, which NVIDIA sized for 24GB. Most Open-Prem deployments will pair a smaller local model for routine work with a larger model, local or via open-weight API, for harder tasks.

Updated deployment tiers

Scale	Hardware	Example models	Notes
Agent fleet / small team	One to four M5 Ultra Mac Studios	Qwen3.8-27B, Muse Glimmer, Granite 4.2, Nemotron 3.5 Bolt	Starts at \$5,499 per unit; 512GB pricing pending
Department (10 to 50 users)	Multi-GPU B200/B300 or MI350 server	GLM-5.3-Flash, MiMo-V2.6-Flash, Nemotron 3.5 Super VL	Check current OEM quotes; prices rose this year
Enterprise (50+ users)	B300 or MI450 clusters; Rubin from 2027	MiMo-V2.6-Pro, GLM-5.3, DeepSeek V4, Kimi K3	Trillion-parameter models need multi-node or very high memory

I have removed the dollar ranges from V3's tiers for the GPU rows. Those ranges were built on prices that have since risen, and I would rather give no number than a stale one. Get current quotes before budgeting.

The Compliance Imperative

The EU deadline moved, but the obligations did not shrink, and breach data got worse. The compliance case for keeping AI systems and data inside the company is stronger than in April.

EU AI Act: high-risk deadline moved to December 2027

The Digital Omnibus was adopted as [Regulation \(EU\) 2026/1744 and entered into force on July 27](#), six days before the original deadline.

Obligation	Original date	Current date
Standalone high-risk systems (Annex III: hiring, credit, education, critical infrastructure)	Aug 2, 2026	Dec 2, 2027
High-risk AI in regulated products (Annex I: medical devices, machinery)	Aug 2, 2027	Aug 2, 2028
Article 50 transparency (chatbot disclosure, synthetic content marking)	Aug 2, 2026	Aug 2, 2026, in force

The Omnibus also [strengthened the AI Office's supervisory powers and added two prohibited practices](#). The extra 16 months gives organizations time to build documentation and conformity processes. On-premises deployment still gives the clearest audit trail: the organization can show exactly which model version ran, on what data, under which policy.

IBM Cost of a Data Breach 2026

IBM's [2026 report, released July 29](#), set new records.

Measure	2025 report (cited in V3)	2026 report
Global average breach cost	\$4.44M	\$4.99M
US average breach cost	\$10.22M	\$11.5M
Breaches involving shadow AI	20%	43%
Average cost of a shadow AI breach	\$4.63M	\$5.39M
Organizations lacking AI governance	63%	68%
Time to identify and contain	Falling	247 days, first increase in five years

IBM also found that [more than one in four organizations experienced an AI-driven attack, which added about \\$1 million per incident](#), and that [21% of shadow AI incidents led to a regulatory fine](#).

The shadow AI finding now has a parallel in software. When employees can generate working applications with coding agents, ungoverned internal software becomes a new source of exposure. The answer is the same in both cases: give people approved tools that run on infrastructure the company controls, so the sanctioned path is easier than the unsanctioned one.

GDPR

Cumulative GDPR fines reached about [EUR 6.11 billion across 2,685 recorded fines](#) as of March 1, 2026, according to the CMS Enforcement Tracker Report. The Omnibus also amended the GDPR to create a lawful basis for processing special category data in bias and fairness testing, which affects how AI teams can use sensitive data in evaluation.

Total Cost of Ownership: What You Compare Against Decides the Answer

Self-hosting pays back quickly against proprietary frontier APIs and slowly, if at all, against open-weight APIs. In V3 I used a single blended API price. The spread between proprietary and open-weight pricing is now wide enough that the comparison point changes the conclusion.

Current API prices (per million tokens)

Model	Type	Input	Output
GPT-6 Astra	Proprietary	\$10.00	\$50.00
Kimi K3	Open weights, hosted	\$3.00	\$15.00
Mistral Large 4	Open weights (pending), hosted	\$1.36	\$4.18
Step 5 Preview	Open weights (custom license), hosted	\$1.00	\$2.70
MiMo-V2.6-Pro	Open weights, hosted	\$0.435	\$0.87

Model	Type	Input	Output
MiniMax M3	Open weights (custom license), hosted	\$0.30	\$1.20

Standard-tier list prices for shorter prompts. GPT-6 Astra charges more above 272K input tokens, and MiniMax M3 doubles its rates above 512K.

A worked payback example

Assume a \$40,000 hardware purchase, 20 million tokens per day, and a 3:1 ratio of input to output tokens.

$$\text{payback (months)} = \frac{\text{hardware cost}}{\text{tokens per day (M)} \times \text{blended price per M} \times 30.4}$$

Compared against	Blended price per M tokens	Monthly API cost	Payback
GPT-6 Astra	\$20.00	\$12,160	3.3 months
V3's blended assumption	\$5.00	\$3,040	13.2 months
MiMo-V2.6-Pro via API	\$0.54	\$331	About 10 years

These figures exclude power, staff time, and maintenance on the self-hosted side, so real payback is longer in every row.

The conclusion for V4: price alone justifies self-hosting when the alternative is a proprietary frontier API at volume. When the alternative is an open-weight model through a hosted API, the case for self-hosting rests on data sovereignty, regulatory requirements, latency, fine-tuning, and protection of internal code and data. Those are often the deciding factors for regulated industries, and the business case should name them directly.

There is a third option. Running the same open-weight model through a hosted API during development, then moving it in-house once volume and compliance needs justify the hardware, avoids lock-in because the weights are the same in both places.

Updated decision framework

Scenario	Recommended approach	Reason
Under 5M tokens/day, no sensitive data	Open-weight model via hosted API	Lowest cost; same weights can move in-house later
Replacing heavy proprietary API use at volume	Self-host	Payback in months against frontier API pricing
Regulated data (HIPAA, GDPR, EU AI Act high-risk)	Self-host	Audit trail and data residency

Scenario	Recommended approach	Reason
Agent fleets running 24/7	Local hardware (M5 Ultra cluster or GPU server) with vetted internal skills	Fixed cost; data stays local
Internal software built with coding agents	Self-hosted coding model, permissive open-source base, closed result	Source code and specifications never leave the network
Bilingual Chinese/English	GLM-5.3-Flash or Qwen3.8	Carried from V3; test on your own data
EU sovereignty requirement	Mistral Large 4 (once weights ship) or Muse Glimmer	European training, or Apache 2.0 with no EU restriction
Need permissive license with no conditions	MiMo-V2.6, GLM-5.3-Flash, DeepSeek V4, Granite 4.2, Muse Glimmer	MIT or Apache 2.0

The Open-Source Creative Stack

The creative stack from V3 still runs on the same hardware, but two license details need correcting and image generation has a new leader in waiting.

Category	V3 pick	October 2026
Video	LTX 2.3	Still Lightricks' current model; a unified LTX Trainer added in June. Some sources list the license as Apache 2.0 and others as the LTX-2 Community License ; confirm before commercial use. NVIDIA's Cosmos 3 (June) ranked first among open models on text-to-image and image-to-video at launch.
Human video with speech	daVinci-Magi Human	Not re-evaluated for this edition.
Image	FLUX.2, Stable Diffusion 3.5	FLUX 3 announced July 23: one model for images, video with audio, and robotics. Proprietary at launch; Black Forest Labs says open-weight versions will follow later this year.
Music	ACE-Step 1.5	Still current; no successor found.
Speech	Voxtral TTS	The open weights are CC BY-NC 4.0 , non-commercial ; commercial use runs through Mistral's API. For self-hosted commercial speech,

Category	V3 pick	October 2026
		StepFun's Step Audio EditX (Apache 2.0) scores higher on Speech Arena (1,118 vs 1,077).

I described Voxtral TTS in V3 as filling the gap for local speech synthesis without cloud dependency. That is true technically, but not for commercial use of the open weights. I have corrected it here.

The creative stack also connects to the application-layer argument. FilmCraft, PhotoCraft, and LightCraft are open-source editors that agents can drive through MCP. Paired with locally hosted generation models, they form a production pipeline that runs entirely on company hardware, with no per-seat creative software subscription.

Conclusion

The first Open-Prem paper argued that companies could own their AI models, and V3 extended the argument to agents. This edition applies it to application software.

The best open-weight model scores twelve points below the best proprietary model on the Artificial Analysis index, and since June open-weight models have processed most of the tokens on OpenRouter. On coding, Xiaomi's vendor-reported DeepSWE tables put the best open models level with Claude Opus 5 and GPT-6 Astra; no independent comparison against the newest closed models is available yet. Agent frameworks run on local hardware, though their skill registries need internal vetting. And this month, a single developer with a frontier coding model produced alpha-stage open-source reimplementations of Photoshop, Word, and Excel.

The reasons to keep AI in-house have not weakened. Regulators extended the high-risk deadline but kept the obligations. Breach costs and shadow AI exposure rose. GPU prices rose. Hosted open-weight models are cheap enough that self-hosting is now a decision about control more often than a decision about cost.

My expectation for the next year is that enterprises will build more of their own software, starting from permissively licensed open-source bases that coding agents keep producing, and will keep the result inside the company. Not every application will move. Systems of record will stay with vendors for years. McKinsey's survey measured the early part of this: 32% of organizations had declined at least one software product or feature because they could build it in-house with agentic coding tools.

About the Author

David Borish is an AI Strategist with 25 years of experience in technology and innovation. He is a keynote speaker on emerging technologies and a guest lecturer at NYU, and mentors AI practitioners through the NYU Stern Frontier Labs Program. He developed the Open-Prem Inflection Point framework. His analysis has appeared in Forbes, TechCrunch, and The Canadian Press. He is the author of the forthcoming book, *The Tony Hawk Paradox*

[DavidBorish.com](#) · [TheAISpectator.com](#)

Sources

Leaderboards and models

- [Artificial Analysis LLM leaderboard](#) and [Intelligence Index v4.3](#)
- Artificial Analysis model reports: [Reflection Beam](#), [Nemotron 3.5 Super VL and Bolt](#), [Nemotron 3 Ultra](#), [Muse Glimmer](#)
- Xiaomi MiMo-V2.6: [Unite.AI](#), [Let's Data Science](#)
- Z.ai GLM-5.3: [IntuitionLabs](#), [GIGAZINE](#), [Let's Data Science](#)
- Moonshot Kimi K3: [NYU Shanghai RITS](#), [LLM Gateway pricing](#)
- StepFun Step 5: [Hugging Face repository](#), [Forkast on the license](#), [OpenRouter pricing](#)
- Alibaba Qwen3.8: [eWeek](#), [Silicon UK](#), [Let's Data Science](#)
- DeepSeek: [Hugging Face model card](#), [MorphLLM](#), [Build Fast with AI](#), [The Next Web on V4.1-Flash](#)
- MiniMax: [M3 on Hugging Face](#), [DataNorth](#), [Puter pricing guide](#)
- Meta: [Futurum on Muse Glimmer](#), [The New Stack on Muse Spark Open](#)
- NVIDIA: [The New Stack on Nemotron 3.5 Bolt](#), [MarkTechPost on Nemotron 3.5 Super VL](#)
- Mistral Large 4: [gradually.ai](#), [Vals AI](#), [LLM Gateway pricing](#)
- IBM Granite 4.2: [DataNorth](#)
- [OpenAI API pricing](#)

Usage

- [OpenRouter rankings](#), usage data through October 15, 2026
- [Capital & Compute on Dirac's compilation of OpenRouter daily usage](#)
- [Vercel AI Gateway Production Index, July 2026](#)

Build era

- [PhotoCraft](#), [WordCraft](#), [GridCraft](#), and [DeckCraft](#) repositories
- [TechSpot: Adobe clones](#) and [TechSpot: Office clones](#)
- [Plagiarism Today: Copyright and the Vibe-Coded Adobe Clones](#)
- [PYMNTS: Adobe settlement](#)
- [AI2 Work summary of McKinsey's State of AI 2026](#)
- [Retool 2026 Build vs. Buy report](#)
- [Forbes on GoDaddy](#), [Business Insider on West Monroe and other in-house builds](#)
- [Lynton on SaaS replacement maintenance](#)

Agents

- OpenClaw releases: [v2026.3.28](#), [v2026.4.7](#); [Outlook Business on Steinberger joining OpenAI](#)
- [ClawHub skills analysis](#), [InsiderLLM on the Koi Security and RankClaw audits](#), [Tencent Zhuque Lab scan](#)
- [OWASP Agentic Skills Top 10](#) and its [repository with the Snyk figures](#)
- NemoClaw: [release notes](#), [version notes](#)

Hardware

- [Apple Mac Studio M5 announcement](#)
- [Digital Citizen on Vera Rubin timing and GPU rental prices](#)

- [Digital Citizen on AMD MI450 sampling](#)
- Vera Rubin: [Spheron on availability](#), [Spheron spec summary](#)
- [VRLA Tech on B300 lead times](#)
- [Capital and Compute on RTX 5090 street pricing](#)

Compliance

- [Cloud Security Alliance on the Digital Omnibus](#), [National Law Review](#)
- IBM Cost of a Data Breach 2026: [Baker Donelson summary](#), [Sentra on shadow AI](#)
- [CMS GDPR Enforcement Tracker Report 2026](#)

Creative

- LTX 2.3: [AI Films on the LTX Trainer](#), [immers.cloud on the license](#)
- DataNorth: [Cosmos 3](#), [FLUX 3](#), [Voxtral TTS license](#)
- [Pinggy on open-source TTS rankings](#)