ACTIVANT RESEARCH

Inference is Accelerating

As the software layer matures, the profit pool moves with it

August 2026

AUTHORS:

Steve Sarracino
Founder and Partner

Jono Vickery
Vice President, Research

About Activant

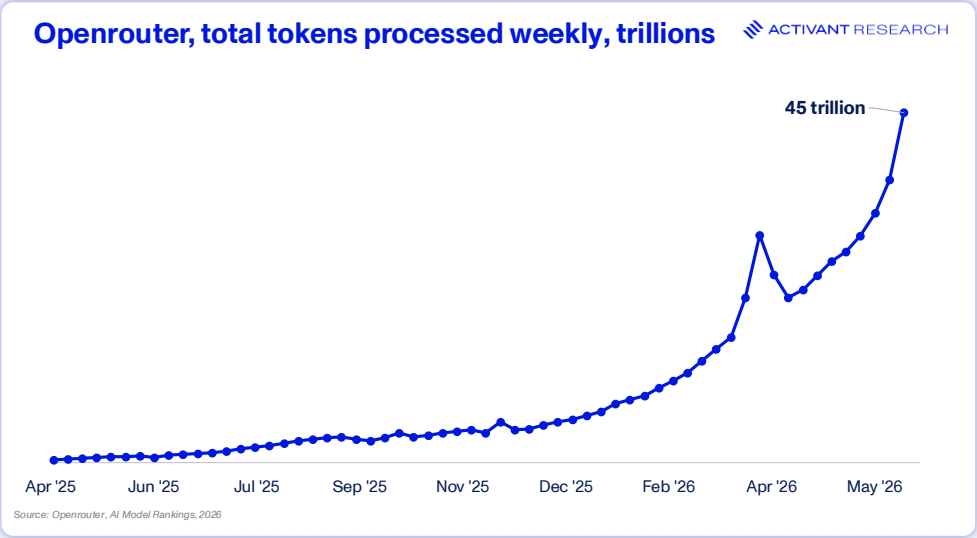
Activant is a research-led global investment firm that partners with high-growth companies. Since 2015, we have invested in category-defining businesses during their most critical phases of growth, partnering with founders who have won their initial battles and are ready for the next challenge.

Our approach pairs deep, proprietary research with patient, flexible capital and hands-on operational partnership. We work alongside founders and leadership teams to refine strategy, strengthen operations, and accelerate sustainable growth.

Activant Research is dedicated to uncovering the most exciting emerging technologies, sectors, and companies we believe will shape the future. Our research-driven perspective informs everything we do, helping us invest at meaningful inflection points and support founders in building enduring, category-leading businesses.

You can find out more about Activant and our research at https://activantcapital.com/.

Amazon Web Services served more tokens in the first quarter of 2026 than in its entire prior history combined.[1] Anthropic's revenue has gone close to vertical, adding the annual revenue of ServiceNow and Workday combined in just Q1, when it went from $9 billion to $30 billion.[2,3] On OpenRouter, every month now eclipses the last. Together AI and Fireworks, both serverless inference providers, have crossed $1 billion in ARR.[4,5] Inference now accounts for roughly two thirds of all AI compute, up from one third in 2023, and CoreWeave confirmed this quarter that more than half of its fleet now serves inference rather than training.[6] **Inference is accelerating.**



In January, we argued that the market was overly focused on training, as inference would ultimately become the dominant workload and software-native providers would lead the infrastructure underlying it. Now it's consensus: inference has to be the center of gravity for AI.

That makes the question of who wins all the more interesting. In this report, we examine where value accrues across the inference stack. Today, there is major signal that OpenAI and Anthropic will benefit from winner-takes-all dynamics. However, we address the flaws in that signal and walk through the profit pools for the infrastructure providers supporting cost-optimized, enterprise-customized tokens. The reality: every layer of the inference stack is commoditizing fast, which sets up the race to be the end-to-end provider that governs the whole stack.

Does Value Concentrate at the Frontier?

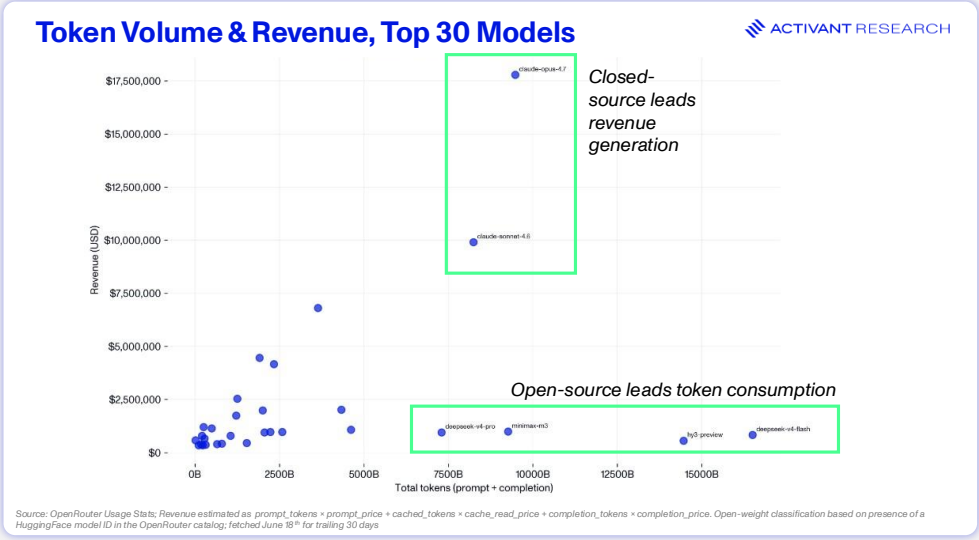
Open source has consistently closed the capability gap with closed source. GPQA Diamond, which we previously used to track open vs. closed model performance, is now largely saturated. Claude Mythos Preview leads at 94.6%, with Gemini 3.1 Pro at 94.3%, and the best open source model, Kimi K3, sits at 93.5%.[7,8] At scores exceeding 90% across the board, the benchmark has stopped reflecting

meaningful differences in practical performance. On the Artificial Analysis Intelligence Index, GLM 5.2, the current open source leader, scores roughly where GPT 5.2 sat six months ago, leaving the six-month lag thesis intact.[9]

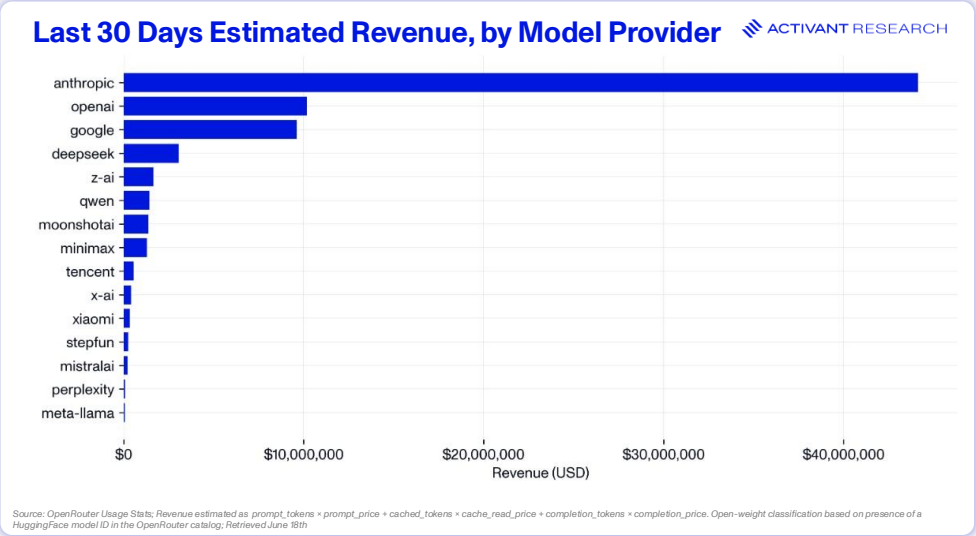
The cost-plus-customization case is running in production. In our compound AI systems hypothesis, we argued that mature deployments would route to open models where fine-tuning and quantization delivered the better outcome: cost efficiency, model control, and compounding improvement.

Fireworks AI now catalogs DeepSeek V4-Pro, Kimi K2.6, and GLM 5.1 in production with fine-tuning supported on models up to 1 trillion parameters. Customers including Cursor, Perplexity, Notion, and Shopify are using fine-tuned models, with Notion reporting latency cut from two seconds to 350ms.[10] Baseten's OpenEvidence deployment runs billions of fine-tuned LLM calls per week across every major US healthcare facility. It drives model latency down over 50% and cost per million characters down 44%.[11] The clearest evidence that open source has cleared the bar for regulated, production-grade workloads is the migration pattern Baseten reports: customers leaving closed APIs, citing rate limiting, pricing, and the inability to own weights or audit model changes.

On OpenRouter, token volume confirms the adoption. In the last 30 days, minimax-m3 processed close to as many tokens as Claude Opus 4.7. Tencent's Hy3 and DeepSeek V4 Flash led the board: the top-volume models on the platform are open source.[12]



Token volume and revenue, however, are not the same. Revenue is concentrating sharply in closed source, and the divergence is more extreme than the capability benchmarks suggest. Performance leadership at the frontier still commands a premium that open source currently cannot capture.



Which opens up the question: does the inference profit pool concentrate at the Frontier Labs?

Observable Data Shows: We're still in the Tokenmaxxing Era

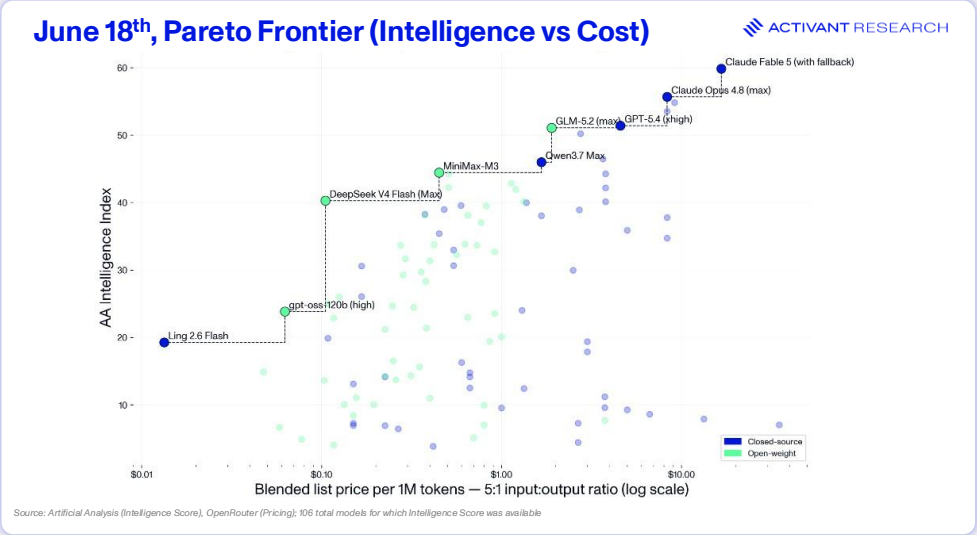
Tokenmaxxing is the mode of deployment that emerges when teams, uncertain which model fits a task, default to the most powerful one available and run it at maximum context, with little regard for cost. The scale of what this looks like in practice has become genuinely striking. Meta employees consumed 73.7 trillion tokens inside 30 days on an internal leaderboard called "Claudeonomics," with internal AI costs approaching billions for the year.[13] One unnamed enterprise ran up a $500 million Claude bill in a single month after failing to set usage limits on employee licenses.[14] Sam Altman has acknowledged the pattern is now "kind of a meme": companies reporting they burned through their entire 2026 AI budget in Q1 and asking OpenAI to make its models cheaper.[15]

That's the environment where we're trying to find signal on model usage.

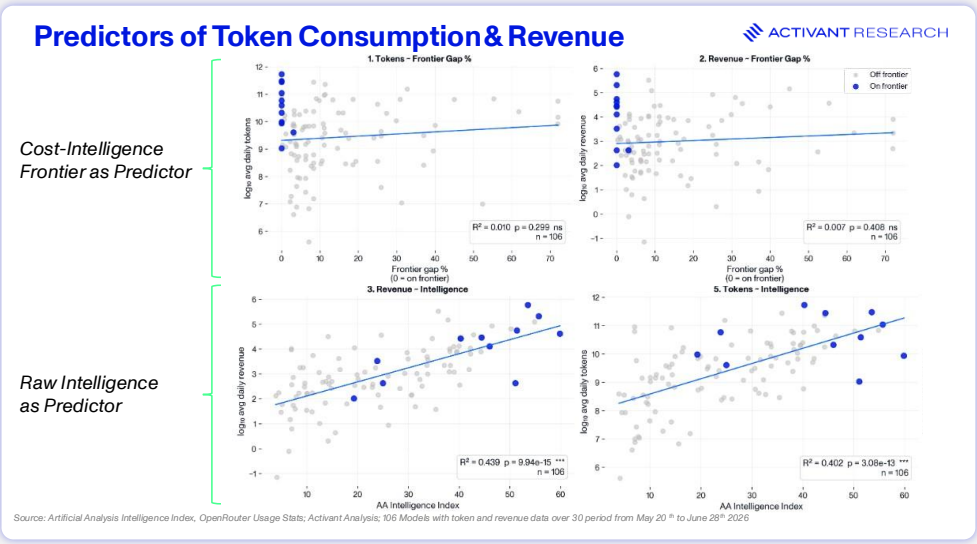
There are early signs that unconstrained AI spend is cooling. Uber burned through its full-year AI coding budget by April and then capped individual engineers at $1,500 a month per tool, with pre-cap bills running $500 to $2,000 per engineer per month.[16] Microsoft canceled most of its internal Claude Code licenses at financial year-end, citing token costs, and redirected thousands of employees to GitHub Copilot CLI.[17] These corrections are real, but they are corrections at the margin. The most diagnostic evidence points the other way.

The strongest signal comes from where dollars actually accrue. In an efficient market, spending should route to the Pareto-optimal model for each task: the point on the cost-intelligence curve where no cheaper model can perform the job as well. Email summarization goes to GPT-OSS 120B. Drug discovery goes to

Claude Fable 5, where cost does not enter the decision. The Pareto frontier is the set of models where that trade-off is correctly calibrated; a model sitting off the frontier is strictly dominated, meaning a cheaper and smarter alternative exists.

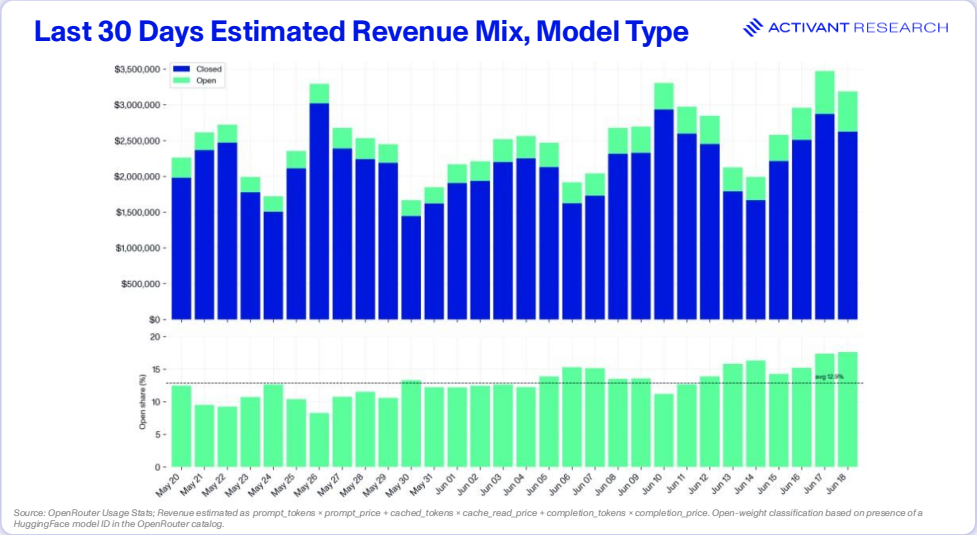


The data suggests none of this calibration is happening. Proximity to the Pareto frontier has zero predictive power over token consumption. Sonnet 4.6 was the fifth most-used model in the dataset while sitting significantly off the frontier, meaning users are routing to it despite cheaper and smarter alternatives being available.[18] The only variable with meaningful predictive power is raw intelligence ranking. OpenRouter users are not, on average, optimizing for cost efficiency. They are optimizing for the smartest model they can access.



Models are not being treated as commodities. If they were, token consumption would track the Pareto frontier. Instead, it tracks raw intelligence rankings, with brand and familiarity reinforcing the bias. We are still in the Tokenmaxxing era.

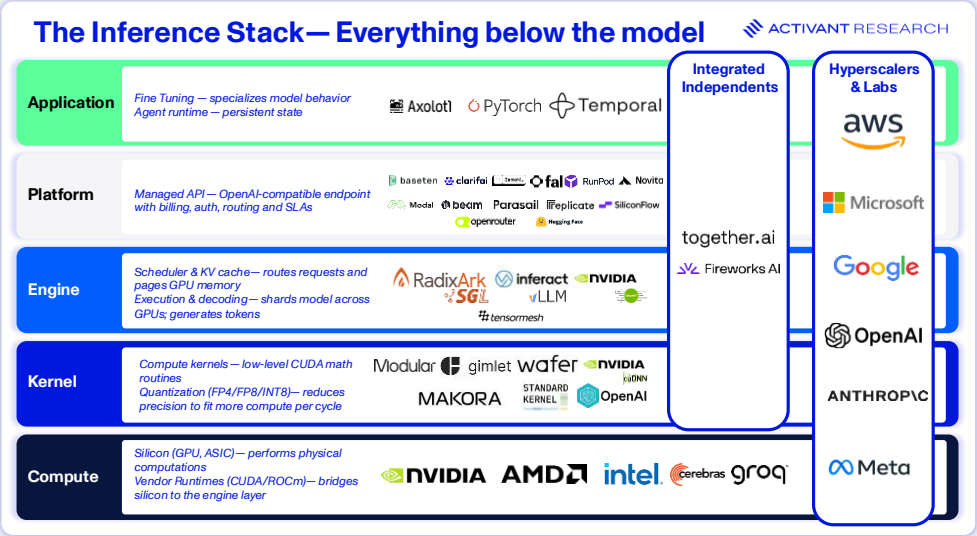
And the reality is that it's tough to draw any meaningful long-term conclusions about where inference profit pools will land while the present data is conflated by this environment.



The caveat is that our use of OpenRouter data substantially overstates how far this extends. It is a single aggregator, capturing just 5–10% of global AI token volume; Fireworks says it alone processes 25 trillion tokens a day, roughly nine times OpenRouter's daily volume.[19] The platform skews toward startups and early-stage experimentation. Enterprise deployments concentrate on Hyperscaler APIs, where compliance, data sovereignty, and cost discipline drive the buying decision. Those workloads are exactly the ones that we expect to shift towards open source but are absent from the signal. Finally, OpenRouter does not host fine-tuned models, which make up 95% of Fireworks' token volume, and that is precisely where open source beats closed source on both performance and cost.

We expect the Tokenmaxxing era to pass, and for many enterprises' procurement cycles, it likely never began. In our view, closed-source frontier labs will capture a large share of the inference profit pool, but less than the OpenRouter snapshot suggests.

The battle for the remainder is what's most interesting, and with the proliferation of open source models, we believe that the profits that don't accrue to OpenAI & Anthropic, accrue to infrastructure providers. Those are the inference clouds, Hyperscalers and neoclouds that we wrote about in our AI Infrastructure series, and the events of the last 6 months make the whole market worth revisiting.



Engines Get Capitalized

In AI Infra Compute 4-4 - Serverless Inference, we argued that the technical depth underneath the API endpoint, everything from a custom filesystem down to the kernel running the math on the GPU itselfwould determine how the inference profit pools get split. It's the deep engineering that a few companies are doing that allows them to solve cold starts; provide better throughput and latency than competitors and protect their gross margins while doing so. But with so many layers under the model, which matter?

A key risk we laid out to that thesis was the existence of open source inference engines like vLLM and SGLang. These tools handle many of the highly complex *memory management* functions that make inference performant and lower the barrier for anyone to compete with companies like Modal, Fireworks and Together AI who are doing that deep, custom engineering.

It's with that in mind that what may have been the key event of the year happened in January, in two announcements days apart. The creators of vLLM and SGLang both incorporated and raised. Inferact (the vLLM founders, with Databricks co-founder Ion Stoica) and RadixArk (the SGLang founders) raised $150M and $100M respectively.[20] Their open-source engines each claim roughly 400,000 GPUs in production (~2% of global GPU supply).[21,22]

vLLM's core innovation, PagedAttention meant that GPUs could allocate memory like a PC operating system, in small, equal-sized blocks that can near-perfectly utilize the GPU's memory. That's compared to just allocating a whole user chat to a block of memory, leaving large portions of GPU memory wasted. vLLM improves the throughput of popular LLMs by 2x to 4x.[23]

SGLang's equivalent breakthrough, RadixAttention, stops different requests from redoing the same work. It keeps already-processed prefixes, like shared system prompts or chat history, in a radix tree to save the system from duplicating

computation. The longer and more repetitive the context, the bigger the saving, which is exactly the shape of agentic and reasoning workloads. The original paper clocked up to 6.4x higher throughput on workloads with heavy prefix sharing.[24]

The point those two innovations make together is that there is no single best engine. The benchmarks bear this out: Nvidia's TensorRT-LLM tends to win raw throughput and latency on Nvidia silicon, SGLang beats vLLM by roughly 30% on shared-context agentic and RAG traffic, and vLLM holds the lowest time-to-first-token and the simplest operations.[25,26] The right runtime is a function of the workload's shape, the hardware underneath it, and how much operational pain a team will tolerate.

The implication is that winners aren't decided at the engine layer. Vendors simply need to match the right engine to the job. That's why Inferact and RadixArk have so many shared customers, and it's the logic behind Baseten's bet: rather than build a proprietary engine, it hosts the commodity ones (vLLM, SGLang, TensorRT-LLM) on globally distributed GPUs and competes on auto-scaling, fault tolerance, and reliability instead.

The capitalization of Inferact and RadixArk brings greater focus, structure and VC-dollars to the commoditization of the engine layer. Anyone building an inference platform gets access to sophisticated scheduling, KV-cache management, and execution machinery. Engine innovation alone doesn't create differentiation.

Ultimately, it's what these companies don't build that really matters. For the actual compute, they call into kernels they didn't write: Nvidia's cuDNN and TensorRT-LLM, the community's FlashAttention and FlashInfer.

The moat was always one layer deeper down the stack, at the kernel. The capitalization of the engine layer makes that thesis look truer than ever.

The Kernel Moat: Real. But, For How Long?

A kernel is the small program that runs directly on the GPU, instructing its thousands of cores how to perform a specific operation, from matrix multiplication to deciding which words to pay attention to when generating the next token. The kernel layer is where the independents genuinely lead. Fireworks runs FireAttention V4 with FP4 quantization on B200s, highly tuned speculative-decoding draft models, and Multi-LoRA serving that holds many fine-tuned variants in memory on a single base model. Together houses a 15-20 person kernel team, the ATLAS speculative decoder, and the FlashAttention lineage itself through Tri Dao.[27] It's scarce engineering that the long tail of inference hosts simply does not have.

The difficulty has everything to do with where data flows through the chip. The GPU's fastest memory is minuscule, lay the data out wrong and the hardware either spills it into slower memory or leaves thousands of threads queuing for the same memory bank, which slows token generation. In the past, that meant hand-writing C++, where a single kernel could take an hour.[28] That work has to be redone for each new chip, number format, and model architecture. Nvidia's CuTe-DSL moves the work to Python, but it doesn't remove the underlying complexity of programming the GPU. Which is why kernel engineers like Tri Dao are regarded in the same realm as the LLM researchers poached for eight- and nine-figure packages.[29]

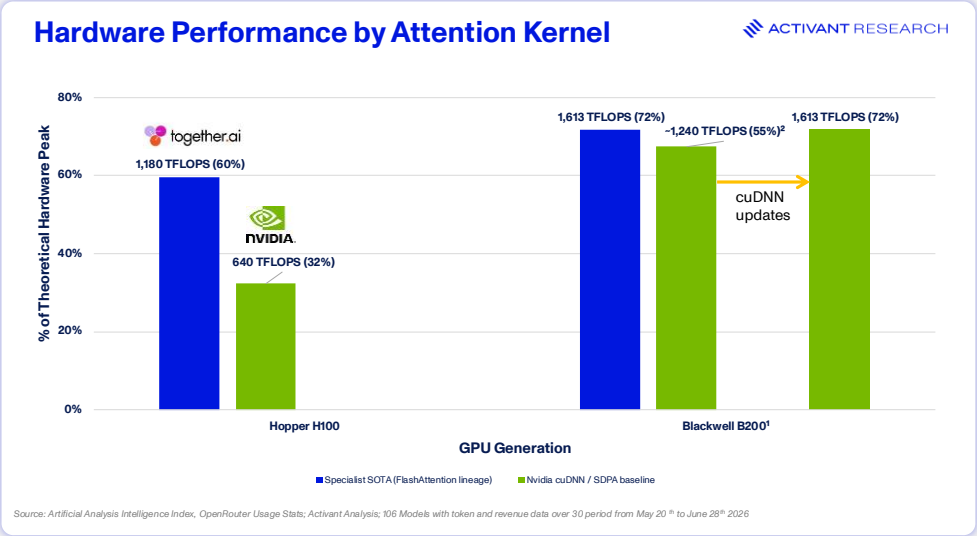
The depth of this engineering work received its strongest possible validation in March, when Microsoft, with effectively unlimited capital and its own silicon program, integrated Fireworks' inference platform into Azure AI Foundry rather than try to rebuild the same thing.[30] Microsoft shipped the integration excluded from its EU Data Boundary commitments, without FedRAMP coverage, and barred from PCI workloads.[31] A company whose enterprise franchise rests on compliance does not compromise it for a 10% speedup.

But there are two tough questions about Kernel engineering as a moat.

How long does the edge last? A kernel is rarely more than a few hundred lines of code, and once it ships, it can be reverse engineered, eroding the edge. We're hearing that kernel engineers are starting to be seen less as an edge and more as an in-house "firefighting team", there to re-optimize the stack in the roughly two-week scramble after each new model architecture drops.[32]

Further, Nvidia closes each gap through cuDNN updates, TensorRT-LLM releases, and NIM packaged microservices, and then the architecture changes and the race restarts. When a new architecture ships, the vendor's own libraries are never optimal on day-zero. The independents have repeatedly proven they reach the new silicon first. On the H100, FlashAttention-3 was extracting 60% of the chip's theoretical peak on long-sequence attention while Nvidia's own cuDNN sat at 32%. That's nearly a 2x gap on the same hardware.[33]

But the gap closes. By the time Together shipped FlashAttention 4 for Blackwell, the gap vs Nvidia's cuDNN was already much smaller (72% vs 55%), and Nvidia quickly folded the same techniques into cuDNN, matching the independent kernel benchmark one for one.[34] You do not bet against Nvidia commoditizing its own complements: every gap it closes defends the CUDA moat.



How long will talent stay scarce? Specifically, what if the labor cost of producing the kernel collapses entirely? Kernels are code, and AI is getting incredibly good at writing code.

Wafer AI used AI agents to do a hardware profile and automate configuration to take the #1 position on Artificial Analysis for AMD hardware (Qwen 3 on MI355X). Traditionally, that would have taken weeks of engineering effort to profile the hardware, write custom kernels, and meticulously tune configurations.[35] Gimlet Labs powers autonomous kernel generation for PyTorch.

Standard Kernel takes it a step further. Acknowledging that automating PyTorch still inherits the trade-offs and inefficiencies inherent in PyTorch, it is building autonomous kernel generation that understands and writes PTX, the low-level assembly language that compiles straight to the GPU. It takes the best methods from high level languages like PyTorch, CUTLASS and Triton, sees their innovations in low-level PTX, and then learns to write at that level. It has already demonstrated performance that beat FlashAttention3, but its results still come from AI and human expert co-design, and that's the key point.[36]

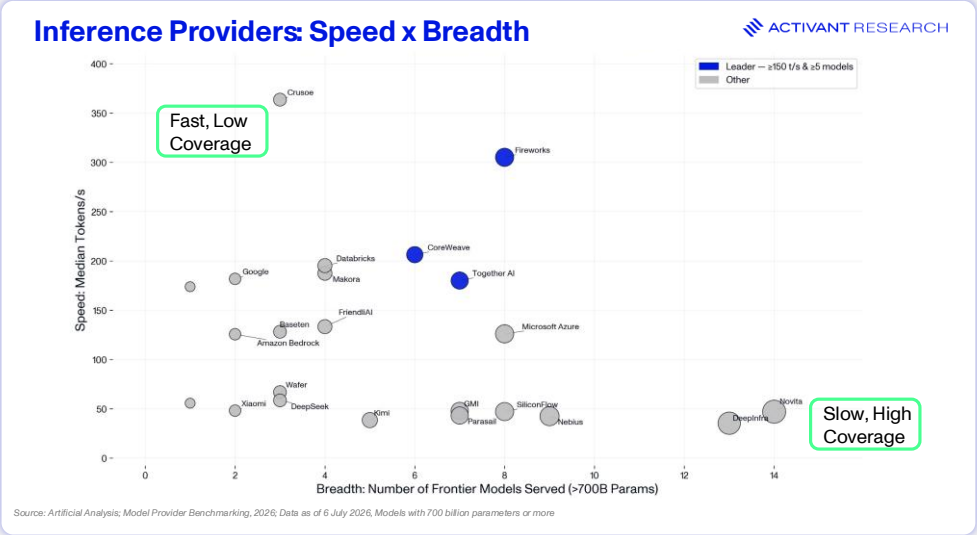
The lack of training data is a critical bottleneck to autonomous kernel generation. Tri Dao, the author of Flash Attention, has noted that we are still "very early". While AI coding tools are improving rapidly, they don't generate correct machine code because the corpus of that data for training is sparse, and largely private. The companies running at this problem will be chasing after scarce kernel talent and asking them to generate AI training data. It's a tough problem but one that will be solved, in the next one to two years according to Tri.[37]

Between Nvidia's tools improving and autonomous kernels maturing, the kernel as a moat will eventually go away, and it's a risk that we didn't call out when we last wrote on the inference providers.

The rebuttal is that the moat was never the kernel alone. The independents' edge comes from a deep, proprietary integration between the kernel and engine as one

system: draft models tuned to the speculative decoder, quantization tuned to the attention kernel, LoRA serving tuned to the scheduler. Open-source stacks are a pile of separately optimized libraries strung together, and that can cost real performance. Our ecosystem feedback put the custom-versus-open gap at ~10% on vanilla, unmodified models, but materially wider on fine-tuned, Multi-LoRA, and latency-exotic workloads.[38] Those are exactly the workloads enterprises pay premiums for and exactly the workloads where the majority of volume for serverless inference clouds lies.[39]

While vLLM and SGLang offer optimized inference engines to the whole market, and Nvidia's cuDNN closes the gap to FlashAttention, Together AI and Fireworks still dominate most open source model provider benchmarks.



Today, it's the integration that matters. But when autonomous kernel generation becomes a reality, the barrier for anyone to optimize and co-design end-to-end inference stack will be significantly lowered. In the long term, inference moats must come from elsewhere.

Performance Wanes, Governance Wins

This is a story we've seen before. Databricks entered the data infrastructure market by offering a managed version of Apache Spark, which led the market in raw performance. Spark was an open source tool for data processing, providing a 10x to 100x improvement on its predecessor, Hadoop.[40] AWS and GCP quickly offered their own versions of managed Spark, which shifted Databricks' advantage into the Lakehouse, an architecture that required deep, proprietary integration between data lakes and data warehouses. Lakehouse offers an obvious total cost of ownership (TCO) advantage over separately managing a data warehouse for BI and a data lake with a processing engine for AI/ML, up to 40% based on ecosystem feedback.[41]

However, Lakehouse or competing architectures are now offered by all three Hyperscalers as well as Dremio, Starburst and Onehouse; with all providers including Databricks supporting Iceberg, the open table format that makes switching providers easy. Databricks responded by shifting its moat for a third time, to **governance**, with Unity Catalog.

As a governance and metadata layer, Unity Catalog makes Databricks the system of record for enterprise data. By centralizing discovery, access control, policy management and spend controls for data, AI assets and applications; Unity Catalog is a data control plane for the enterprise. Today, it's governance, not raw performance that supports Databricks' $134B valuation, $5.4B in ARR growing more than 65%, and net retention above 140%.[42]

The detail that matters for inference: Databricks never tried to out-govern AWS head-on. In 2015 that fight was unwinnable, just as out-complying Bedrock is unwinnable today. Databricks earned the governance layer by first owning the performance advantage, then creating data gravity. That's the framework.

However, there's one critical difference in the AI era, everything moves faster, including the clock on your own advantage. Databricks had roughly seven uncontested years on its compute wedge before managed Spark caught up. Autonomous kernel generation might mean that the inference independents have 18 to 24 months.

Through this lens, the Azure <> Fireworks deal, which so proved Fireworks performance advantage, also shows the moat being mortgaged. Get distribution today, paid for with the customer relationship tomorrow.

From Mortgaging the Moat to Creating Your Own Gravity

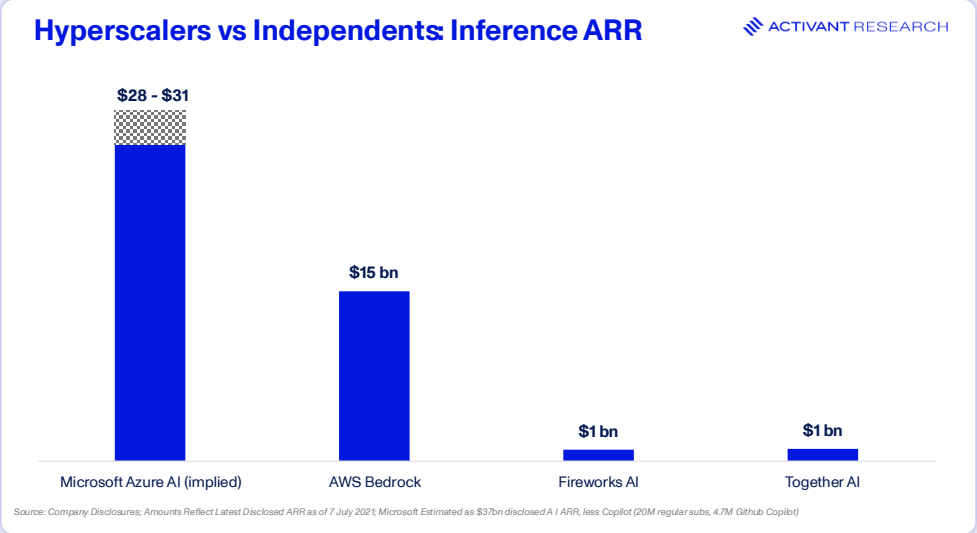
When Fireworks was integrated into Azure, Microsoft kept the billing relationship, the interface, the developer ecosystem, and the governance surface. Fireworks kept the execution layer. The Bring-Your-Own-Weights capability, the deal's retention centerpiece, anchors customers' fine-tuned weights inside Microsoft's Foundry while Fireworks' stack serves them.[43] The gravity lands on Microsoft, which captures all the structural lock-in.

Contrast the precedent directly. Azure Databricks was also first-party Hyperscaler distribution, and it was crucial to Databricks' enterprise scaling. But Databricks kept Unity Catalog, the workspace, and the system of record even inside Microsoft's cloud. Same channel, opposite gravity ledger. The test of a distribution deal is who owns the catalog-equivalent, and on that test the two deals are opposites.

Fireworks didn't sell its moat; it mortgaged it, and Microsoft holds the lien.

The critical issue is that the governance and control layer is exactly where the Hyperscalers win. With their performance catching up, it's the last thing to cede.

Microsoft runs Foundry as a unified control plane: catalog, evaluations, governance, and agent tooling, with the Fireworks integration filling its one performance gap and more than 20 million paid Copilot seats pulling from the application layer above.[44] Amazon's Bedrock reaches 80% of the Fortune 100, with customer spend up 170% quarter over quarter; its IAM and VPC friction is the compliance product, delivering SOC 2, HIPAA, and FedRAMP without custom engineering, while Trainium compresses its cost base underneath.[45] Google runs Vertex on top of TPU v8, whose inference (8t) chip can be networked in a fabric of 9,600 chips, providing substantial inference performance advantages due to pooled memory. 330 Vertex customers processed more than a trillion tokens in the past year, and Vertex Enterprise bundles at a flat $30 per user per month.[46] All three share a structural advantage no independent can replicate: they serve closed and open models side by side, so a customer's model choice never forces a platform choice.



So, the independent path is to create **gravity**: leverage performance as the wedge, create switching costs and become a control plane later.

Fine-Tuning: the Optimization Moat

When a customer fine-tunes a model, the resulting weights are optimized for the memory layout of the engine that produced them. Move them to a rival stack and they will not perform the same, assuming the provider lets you move them at all. Fireworks AI runs the full post-training pipeline: supervised fine-tuning (SFT), which trains a model on labeled examples of the output you want; direct preference optimization (DPO), which trains it on pairs of better and worse answers; and reinforcement fine-tuning (RFT), which optimizes it against a reward function. Running post-training on Fireworks pipeline binds the weights tightly to FireAttention's memory layout, and Fireworks doesn't provide weight export today.

95% of Fireworks' token volume comes from customer models fine-tuned on their proprietary data, which are thus tied to Fireworks' platform.[47]

Consider Cursor, which trained Composer 2, a coding model that beats Anthropic's Opus 4.6 on terminal-agent benchmarks, entirely on Fireworks training infrastructure, before deploying it into production on the same stack.[48]

Together AI also covers post-training including SFT and LoRA. But its RFT path runs on torchforge, the open-source framework it builds with Meta and PyTorch, so the weights stay portable.[49] Baseten is running the opposite play, selling weight ownership and portability as the product, leaving customers in full possession of weights and code. It is a direct bet against the Fireworks thesis, and the demand for it confirms the lock-in is real enough that customers will pay to avoid it. But portability-as-product forgoes the gravity that we argue is where durable margin accrues, and Baseten doesn't optimize kernels, so it competes one rung below the performance wedge. It wins the customers who fear lock-in today; it is not, in our view, the shape of the long-term winner.

Reinforcement Learning: the Co-Location Moat

Reinforcement learning collapses training and inference into a single loop. The model generates outputs (inference), a reward function scores them, and those scores feed the next pass that updates the weights (training). Co-locating them matters due to numerical precision. Training runs at high precision while inference is deliberately run low to go faster, and when the two disagree on the numbers the reward signal corrupts, degrading the run and, at the extreme, diverging it outright.

Fireworks AI now sells this as a service: labs bring their own trainer and run rollout serving, policy updates, and fleet orchestration on Fireworks infrastructure, ensuring identical precision end to end.[50] RadixArk's Miles framework solves the same problem natively for the SGLang stack, which xAI runs both Grok's production inference and its RL post-training on across more than 100,000 GPUs.[51] Modal offers Training Gym, which co-locates training, RL rollouts and isolated environments on Modal infra

In this case, moving to a different stack is theoretically possible, but labs need to re-validate the stability of the new stack from scratch, and the cost of a lost training run could be eight figures, i.e. prohibitive.[52] **The weights can in principle be exported; the loop that trains them cannot.**

Agent Runtimes: the Execution Moat

PocketOS had its entire production database and backups deleted by an agent.[53] Agent runtimes are designed to ensure this never happens again. In this case, the runtime is an isolated, persistent execution environment, or sandbox, where untrusted code runs safely, state survives between steps, and a GPU is on hand when a step needs one. Modal pairs gVisor-isolated code execution with on-demand GPU access and its sandboxes account for more than a third of ARR.[54]

Re-platforming an agent runtime means re-architecting the execution environment and container isolation from scratch, making it that little bit harder to switch from the inference platform that hosts this infrastructure. However, the flaw is that the runtime sits one layer above inference. A customer can keep its sandboxes and agent state on Modal while pointing the model calls at whichever provider is cheapest. But that's kind of what becoming a control plane is all about.

Where Gravity Becomes Control

These moves, across post-training, RL and agent workflows serve to further fuse the customer workflows to the inference providers stack, and once they've built sufficient gravity, the final play is to become a control plane. Become the layer that governs everything above the execution engine: discovery, access control, versioning, evaluation, policy, and spend, across every model, fine-tune, and agent a customer runs, whatever serves them underneath. It's what makes lock-in real after kernel advantages fade.

Independent inference providers are fighting to be the control plane from underneath, with leading execution and switching costs gained through gravity. The hyperscalers build down from governance, already the system of record for enterprise cloud and now fitting high-performance inference underneath. On balance, we believe that the production traction across the three gravity mechanisms bodes well for the independents' position long term.

Capturing The Profit Pool

Inference is accelerating, and many believe that it will be the largest market technology has ever produced. The central question that this article addressed, who benefits most, was of course nuanced. We see the frontier labs as major beneficiaries, but so too the infrastructure providers making open source, custom models possible. However, AI moats erode fast, as we've seen with the inference engine tooling, and expect to see in the next ~24 months on the kernel layer. Thus, infrastructure profits accrue to the provider that converts a performance wedge into gravity, and gravity into a control plane that governs every model, fine-tune, and agent a customer runs.

Critically, as the software tooling that makes AI possible matures, Nvidia's CUDA moat weakens. Soon, AI will write the optimized kernels that made switching to a different hardware provider a non-starter. This is the heterogeneous compute hypothesis: as the software abstracts away from the hardware, the hardware underneath it stops having to be Nvidia's. That doesn't spell the end of Nvidia, but it dents its pricing power and drives profits out of the silicon layer and to the software layer. While we've argued about how the software profit pool gets divided, the larger and more certain shift is that the pool itself is growing and migrating away from the layer that has held the industry’s richest margins for the last three years.

[1] Amazon, Q1 2026 Earnings Call, 2026.

[2] Anthropic, Anthropic expands partnership with Google and Broadcom for multiple gigawatts of next-generation compute, 2026

[3] CapIQ, ServiceNow / Workday Revenue, 2026; Activant analysis.

[4] Fireworks AI, Announcing our Series D and $1B ARR, 2026.

[5] Reuters, Together AI raises $800 million at $8.3 billion valuation, 2026.

[6] Coreweave, Q1 2026 Earnings Call, 2026.

[7] LLM-Stats, LLM Leaderboard, 2026; GPQA Diamond near saturation — Claude Mythos Preview ~94.6%, Gemini 3.1 Pro 94.3%, best open model Kimi K2.6 90.5% (Qwen 3.5 ~88%, DeepSeek V4 close behind).

[8] LLM Gateway, Kimi K3 vs Claude Opus 4.8: Benchmarks, Price, Verdict, 2026

[9] Artificial Analysis, Intelligence Index, 2026.

[10] Fireworks AI, Customer Case Studies, 2026; Notion 2s → 350ms latency, Vercel 40x on a code-fixing model, Quora 3x response-time improvement; named customers per company catalog, May 2026. Vendor-reported case studies.

[11] Baseten, OpenEvidence Customer Case Study, 2026; >50% model latency reduction, 44% cost-per-million-characters reduction, billions of fine-tuned LLM calls per week across every major US healthcare facility. Vendor-reported case study.

[12] OpenRouter / a16z, State of AI: An Empirical 100 Trillion Token Study, December 2025.

[13] The Information, Tokenminimizing: Meta Moves to Curb Employee AI Usage as AI Costs Reach Billions, 2026.

[14] Axios, AI spending, ROI and enterprise costs, May 28, 2026; company unnamed.

[15] Sam Altman, Intelligence at Work (event), 2026; reported by Tom's Hardware. Quote: "My company spent my entire 2026 budget in Q1. Can you make this more efficient?"

[16] Bloomberg, Uber Caps Usage of AI Tools Like Claude Code to Manage Costs, June 2, 2026; capped engineers at $1,500/month per tool after exhausting its full-year AI budget by April; pre-cap spend $500–$2,000/month.

[17] Forbes, Microsoft Ends Claude Code Licenses As It Pushes Copilot CLI, June 1, 2026; most internal Claude Code licences in the Experiences and Devices division cancelled effective June 30, 2026, citing token-based billing, with engineers redirected to GitHub Copilot CLI.

[18] OpenRouter, Model Usage Stats Pages, 2026; Activant Analysis.

[19] Lin Qiao (Fireworks AI), Building a Moat in the Age of AI Agents, Startup Grind, 2026; company claim ("we process more than 25 trillion tokens a day").

[20] Inferact / RadixArk, Funding Announcements, 2026.

[21] Inferact, Investor Disclosures, 2026.

[22] SGLang, Docs, 2026.

[23] Kwon et al., Efficient Memory Management for Large Language Model Serving with PagedAttention, 2023.

[24] Zheng et al., SGLang: Efficient Execution of Structured Language Model Programs (RadixAttention), 2023.

[25] Spheron, vLLM vs TensorRT-LLM vs SGLang: H100 Benchmarks, 2026; Note: secondary technical benchmarks; figures are workload- and config-dependent.

[26] LeetLLM, Choosing an Inference Engine in 2026. vLLM holds the lowest time-to-first-token and is the lowest-effort engine to stand up. Note: secondary technical benchmarks; directional.

[27] Together AI, Inside the Together AI kernels team, 2026.

[28] Activant Capital, Ecosystem Interviews, 2026.

[29] CNBC, Sam Altman Says Meta Tried to Poach OpenAI Staff with $100 Million Bonuses, June 18, 2025; also Fortune, How Much AI Salary, July 11, 2025. Figures disputed by Meta; cited as directional magnitude.

[30] Microsoft / Fireworks AI, Integration Announcement, March 2026.

[31] Microsoft, Foundry Documentation: Fireworks on Foundry Preview Exclusions (EU Data Boundary, FedRAMP, PCI), March 2026.

[32] Activant Capital, Ecosystem Interviews, 2026.

[33] Activant Analysis, Sources: FA3 paper, PyTorch FA3 blog, Lambda FA4 benchmark, FA4 paper.

[34] Tri Dao et al., FlashAttention-4, 2026.

[35] Tensorwave, Wafer Reached #1 Inference Performance for Qwen3.5-397B on AMD, 2026.

[36] Standard Kernel, Announcing Our Seed Round: Is Kernel Generation Solved?, 2026.

[37] Tri Dao, The End of Nvidia's Dominance, Why Inference Costs Fell & The Next 10X in Speed (interview), 2026.

[38] Activant Capital, Expert Interviews, Q2 2026.

[39] Sarah Guo (Baseten Investor), The Untrainable, 2026.

[40] Tech Insider, Spark vs Hadoop 2026: 100x Speed, $134B Bet, 2026.

[41] Activant Capital, Expert Calls, 2026.

[42] Databricks, Series L Announcement and ARR Disclosures, late 2025 / January 2026.

[43] Microsoft, Introducing Fireworks AI on Microsoft Foundry: Bringing high performance, low latency open model inference to Azure, March 2026.

[44] Microsoft, FY26 Earnings Disclosures, 2026.

[45] Amazon, Q1 2026 Earnings Call, 2026.

[46] Google, Welcome to Google Cloud Next, 2026.

[47] Fireworks AI, Announcing our Series D and $1B ARR, 2026; >95% of token volume from customer models fine-tuned on proprietary data.

[48] Sequoia Capital, How Cursor Trained Composer on Fireworks, 2026; Composer 2 trained and served end-to-end on Fireworks infrastructure; beats Anthropic Opus 4.6 on terminal-agent benchmarks per company claims.

[49] Activant Capital, Ecosystem Interviews, 2026.

[50] Fireworks AI, Frontier Lab Training Infrastructure, Now as a Service, 2026.

[51] Accel, Our Investment in RadixArk, May 2026.

[52] Epoch AI, How Much Does It Cost to Train Frontier AI Models?, 2024; frontier training runs now exceed $100M in compute and failed/ablation runs consume 10–30% of that, so a single lost or diverged run runs to eight figures.

[53] The Guardian, Claude AI Deletes Firm Database, April 29, 2026.

[54] Modal, Series C Announcement, May 21, 2026; sandboxes represent more than one-third of ~$300M ARR; 1 billion+ sandboxes launched; 50,000+ concurrent sessions.

Disclaimer: The information contained herein is provided for informational purposes only and should not be construed as investment advice. The opinions, views, forecasts, performance, estimates, etc. expressed herein are subject to change without notice. Certain statements contained herein reflect the subjective views and opinions of Activant. Past performance is not indicative of future results. No representation is made that any investment will or is likely to achieve its objectives. All investments involve risk and may result in loss. This newsletter does not constitute an offer to sell or a solicitation of an offer to buy any security. Activant does not provide tax or legal advice and you are encouraged to seek the advice of a tax or legal professional regarding your individual circumstances.

This content may not under any circumstances be relied upon when making a decision to invest in any fund or investment, including those managed by Activant. Certain information contained in here has been obtained from third-party sources, including from portfolio companies of funds managed by Activant. While taken from sources believed to be reliable, Activant has not independently verified such information and makes no representations about the current or enduring accuracy of the information or its appropriateness for a given situation.

Activant does not solicit or make its services available to the public. The content provided herein may include information regarding past and/or present portfolio companies or investments managed by Activant, its affiliates and/or personnel. References to specific companies are for illustrative purposes only and do not necessarily reflect Activant investments. It should not be assumed that investments made in the future will have similar characteristics. Please see "full list of investments" at activantcapital.com/companies/ for a full list of investments. Any portfolio companies discussed herein should not be assumed to have been profitable. Certain information herein constitutes "forward-looking statements." All forward-looking statements represent only the intent and belief of Activant as of the date such statements were made. None of Activant or any of its affiliates (i) assumes any responsibility for the accuracy and completeness of any forward-looking statements or (ii) undertakes any obligation to disseminate any updates or revisions to any forward-looking statement contained herein to reflect any change in their expectation with regard thereto or any change in events, conditions or circumstances on which any such statement is based. Due to various risks and uncertainties, actual events or results may differ materially from those reflected or contemplated in such forward-looking statements.