runpod

The State of AI

March 2026

Table of Contents

The State of AI

The most widely deployed open-source LLM isn't Llama anymore.

Upscaling now outpaces video generation.

Nearly 70% of image workflows run through a single tool.

These aren't predictions or benchmarks. They're signals from production.

At Runpod, we power AI workloads for more than 500,000 developers and companies worldwide. That scale, combined with our strong data foundation, gives us a rare view into how AI is actually being used: not what's trending on X, but what's running in production, hour after hour.

We analyzed anonymized serverless traffic, GPU utilization, and user behavior across our platform to surface the infrastructure patterns that define AI in 2026. Here's what we're seeing on the ground.

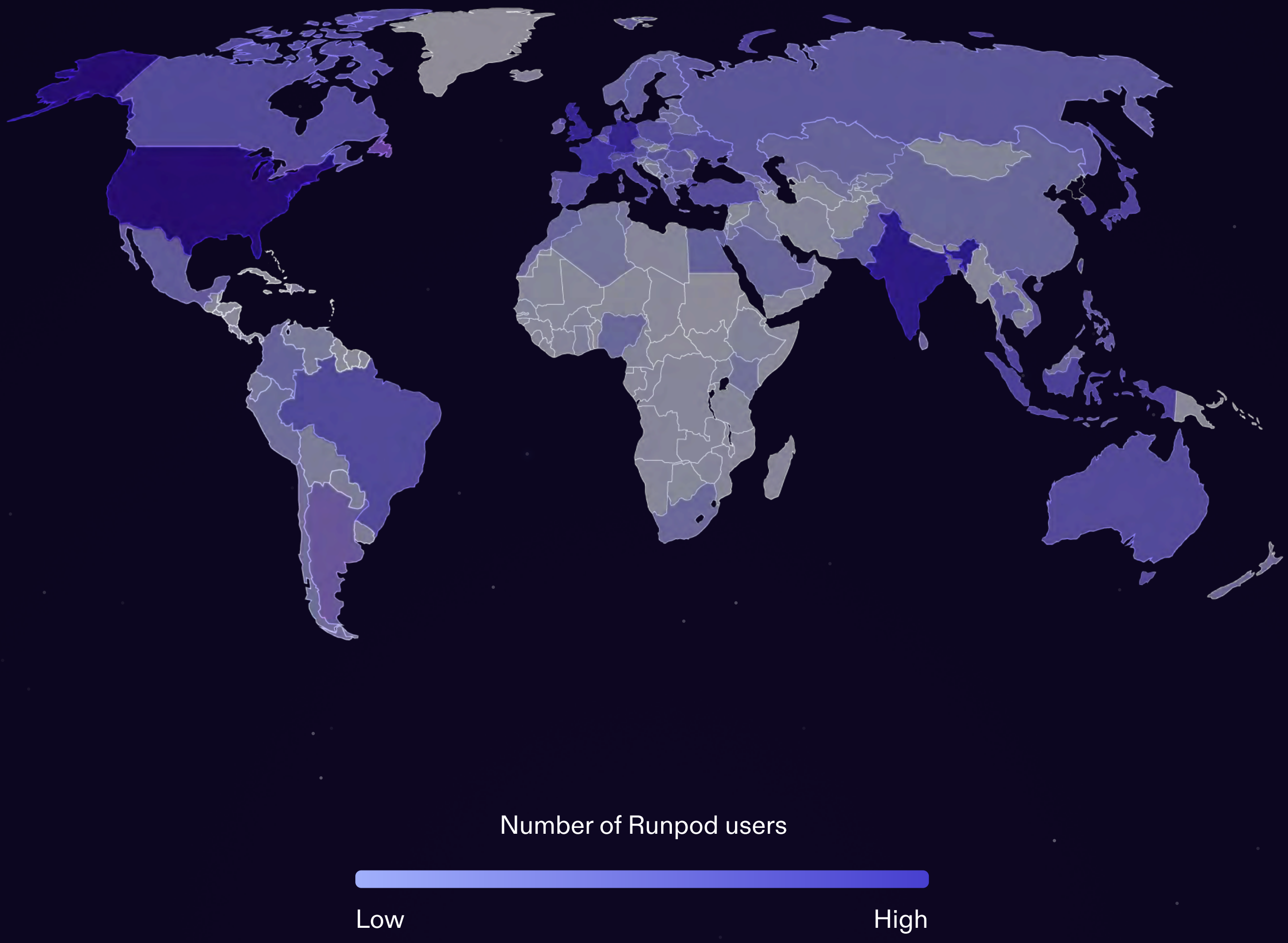
Who is Building with AI?

> Runpod's user base spans 183 countries, covering every inhabited continent, with the United States leading at over a quarter of all users.

India follows as the second-largest market at 7% of users, reflecting strong demand for AI infrastructure in rapidly growing developer ecosystems. Europe collectively represents nearly a third of the platform's user base, with Germany, the United Kingdom, the Netherlands, Poland, Spain, and Italy among the top markets. Asia-Pacific continues to grow, with South Korea, Japan, Vietnam, Thailand, and Singapore all showing meaningful adoption. Brazil and Canada round out the top markets in the Americas. This global distribution underscores that demand for GPU compute isn't limited to Silicon Valley. Developers and companies worldwide are building AI applications and need accessible, flexible infrastructure to do it.

183
countries

Runpod Users by Country



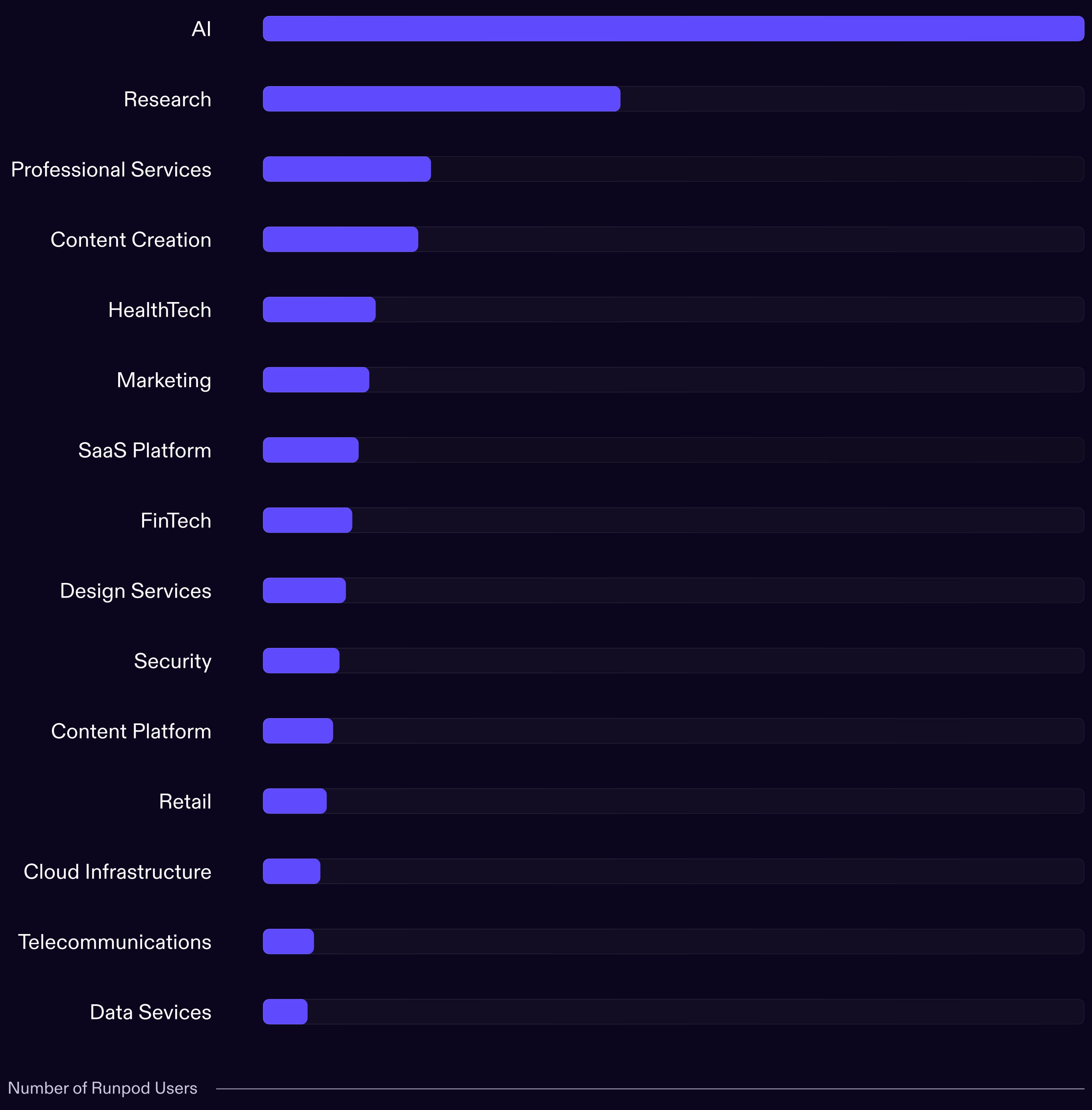
Only a third of Runpod's base is classified as AI-native, which highlights how widely the impact and value of AI has been recognized and adopted. The AI industry is far from monolithic, encompassing a wide range of subcategories, from AI tooling to foundational model development. Following AI-native, research represents the next largest share of users. This includes universities and research labs driving experimentation and cutting edge innovation, from advancing research in LLM post-training and reasoning to protein structure prediction and drug discovery workflows. Beyond these segments, Runpod serves a diverse mix of industries: professional services, content creation, marketing, design services, SaaS platforms, and security all show meaningful adoption. HealthTech and FinTech define the top enterprise verticals.

> GPU compute is becoming essential infrastructure across the economy, not just for AI-native startups.

2/3

Outside AI

Top Industries



Stable Diffusion Leads Image Generation, But the Field Is Shifting

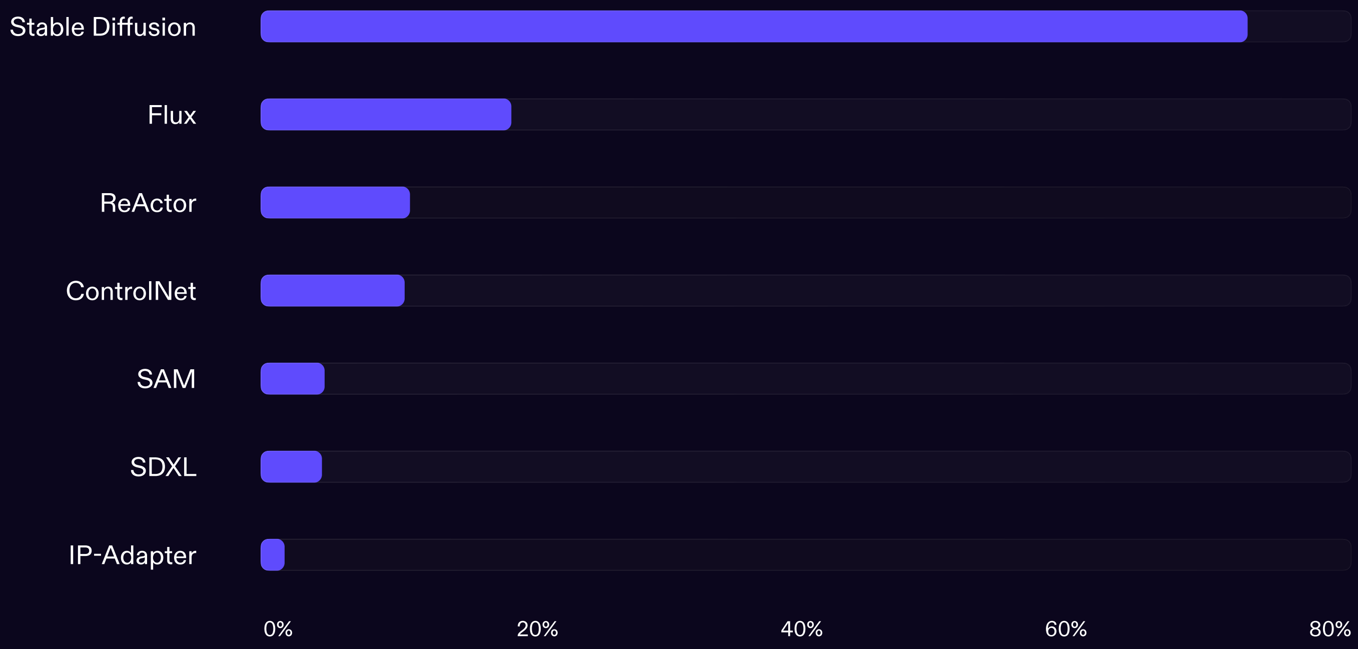
> Stable Diffusion remains the dominant force in image generation, powering 67% of image endpoints.

But Flux has emerged as a serious challenger. Despite being newer to the scene, Flux endpoint creation nearly quadrupled in the second half of 2025, outpacing Stable Diffusion's growth over the same period. This trajectory suggests it could compete for Stable Diffusion's crown. Meanwhile, in contrast with earlier versions of Stable Diffusion, Stable Diffusion 3 has struggled to gain traction, with only a handful of endpoints using it. For practitioners, betting on Stable Diffusion remains safe, but Flux's rise suggests it deserves serious attention. We expect Flux will continue gaining market share, potentially capturing over one third of image endpoints by mid-2026.

67%

of image endpoints powered by Stable Diffusion

Top Image Models



Share of Image Model Usage (by number of endpoints)

In terms of tooling, ComfyUI has become the de facto standard for image generation, powering over two-thirds of image endpoints with its node-based approach. This dominance reflects a broader shift toward modular, customizable pipelines rather than simple text-to-image calls. If you're building image generation workflows, investing in ComfyUI expertise is increasingly essential. It's where the ecosystem has converged.

We're also seeing image generation intersect with language models. A meaningful share of image endpoints now incorporate text operations, most commonly using Qwen or GPT-4o. We’re seeing a variety of use cases: auto-generating prompts from reference images, building searchable visual databases, and creating high-quality descriptions for training data. If you're not yet incorporating LLMs into your image pipelines, you may be missing efficiency gains that leading practitioners have already adopted.

2/3

of image endpoints using ComfyUI

Wan Dominates Video Generation, and Post-Processing Is Baked Into Most Pipelines

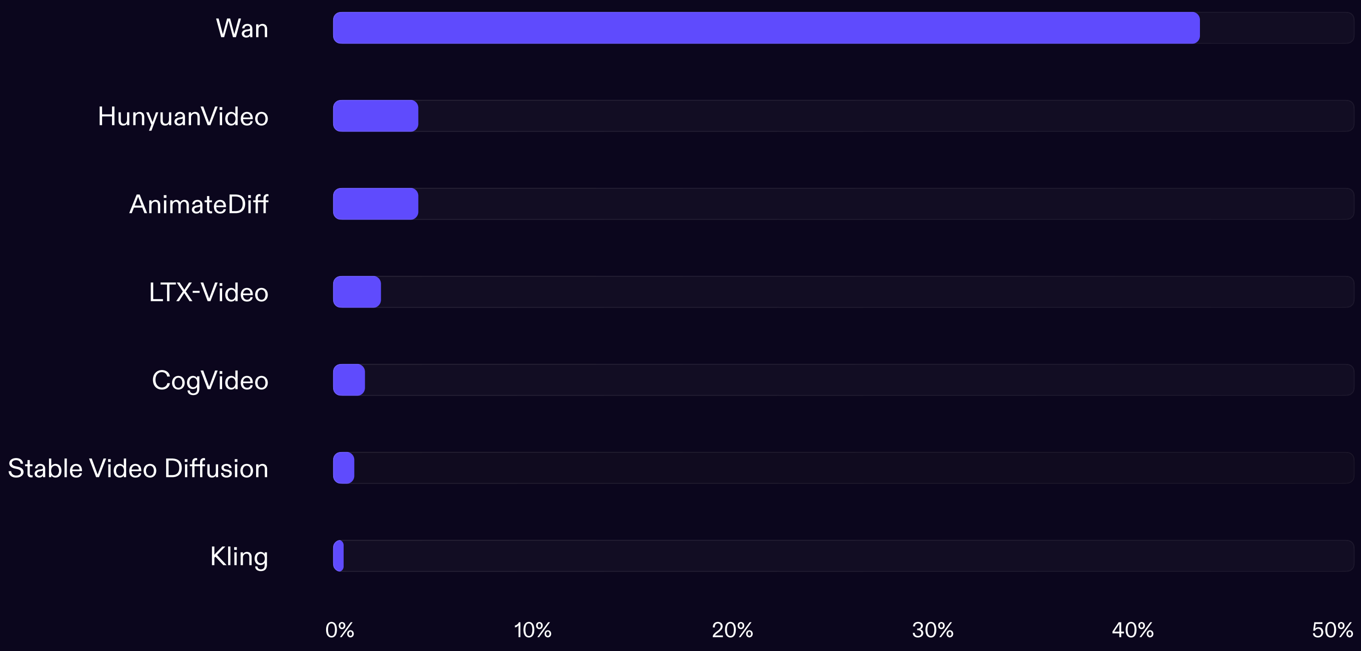
> Wan has essentially captured the open-source video generation market, powering 43% of Serverless endpoints doing video generation.

Proprietary models like Veo 3 and ByteDance's Seedance are competitive, but Wan is the natural choice for developers who want control over their infrastructure and costs. ComfyUI's dominance is even more pronounced for video than for image, with video workflows running almost exclusively through its node-based interface. This makes sense given the complexity of video pipelines. ComfyUI's modular approach excels at chaining together generation, upscaling, and other post-processing steps.

43%
of video generation endpoints powered by Wan

LTX is small but growing fast. Wan currently leads, but LTX has grown over fifteen-fold since November, with a notable spike following the LTX-2 release in early January 2026. Wan has more than doubled over the same period. Healthy growth, but not the explosive trajectory we're seeing from LTX.

Top Video Generation Models



Share of Video Generation Model Usage (by number of endpoints)

But the bigger insight is what users are actually doing with their pipelines. 70% of video generation endpoints also perform video upscaling or enhancement, revealing that post-processing is a critical part of real-world workflows. The typical pattern involves generating lower-resolution videos first, then upscaling the best results to final quality. This reflects a "rolling the dice" strategy: users maximize their creative options by running many fast, lower-quality generations, cherry-picking the winners, then investing compute only in upscaling those. We expect video workflows will remain "draft then refine" for the foreseeable future; real-time generation is likely several years out for most production use cases.

It's also worth noting that image-to-video is more common than text-to-video. Practitioners are largely starting with a reference image and animating from there rather than generating from text prompts alone. This approach gives creators more control over the final result: a reference image locks in subject identity, style, and composition before any frames are generated.

70%

of video generation endpoints also doing upscaling or enhancement

Whisper Leads Transcription, While Text-to-Speech Remains Fragmented

> Audio workflows look fundamentally different from image and video. While those domains are primarily about creation, audio is consumption-first: transcription is six times more common than text-to-speech.

This makes sense given the sheer volume of audio the internet produces daily, from podcasts to interviews to calls. Transcription acts as a multiplier on existing raw audio content, making it searchable, translatable, and accessible to downstream AI workflows. Whisper has become the clear winner, powering over two-thirds of audio endpoints. Its dominance extends beyond standalone use; there's meaningful overlap with video workflows, where Whisper handles the audio extraction and captioning layer. If you're building media pipelines, Whisper integration is likely already on your roadmap.

2/3
of audio endpoints powered by Whisper

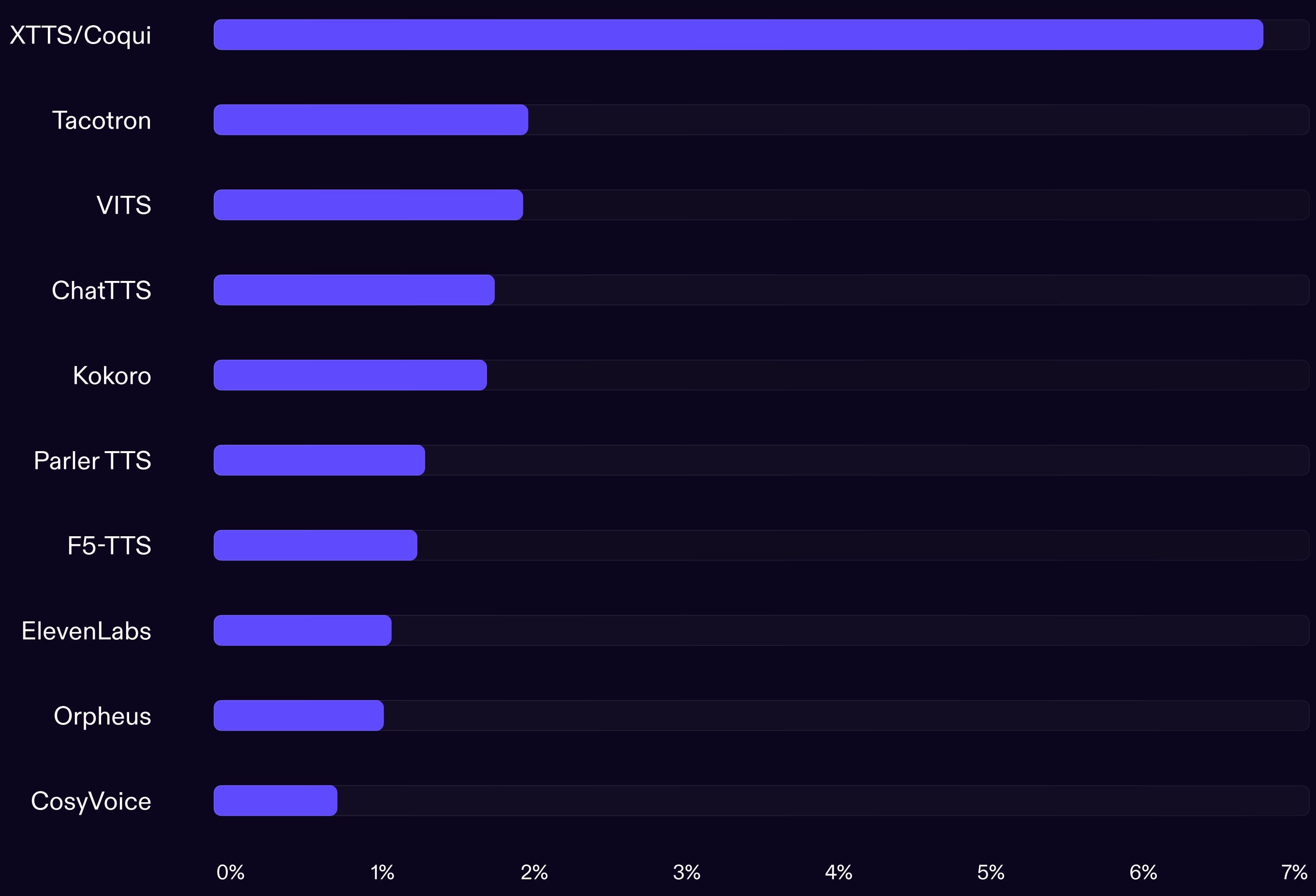
The technical architecture differs too. Unlike image and video workflows that have converged on ComfyUI's node-based approach, audio workflows remain Python-native and PyTorch-based. This reflects both the relative maturity of audio tooling and the simpler, more linear nature of most audio pipelines. Transcription doesn't require the complex branching and post-processing chains that make node-based interfaces so valuable for visual media.

On the generation side, the text-to-speech market remains fragmented. No single model has achieved dominance: XTTS, VibeVoice, ChatterboxTTS, and Minimax Speech are all in the mix, but none has broken away from the pack. The leading model (XTTS) has only 7% of endpoints. This fragmentation suggests the space is still early and ripe for disruption.

x6

Transcription vs. TTS

Top Text-to-Speech Models



Share of Text-to-Speech Model Usage (by number of endpoints)

Qwen Wins the Open-Source LLM Race

> **Qwen has emerged as the dominant self-hosted LLM on Runpod, outpacing Meta's Llama family to effectively win the open-source model race.**

For pure text workloads, Qwen leads, followed by Llama and GPT-4o. This is a notable shift. While Llama dominated open-source mindshare for much of the past two years, Qwen's combination of performance and efficiency has won over practitioners for deploying models at scale.

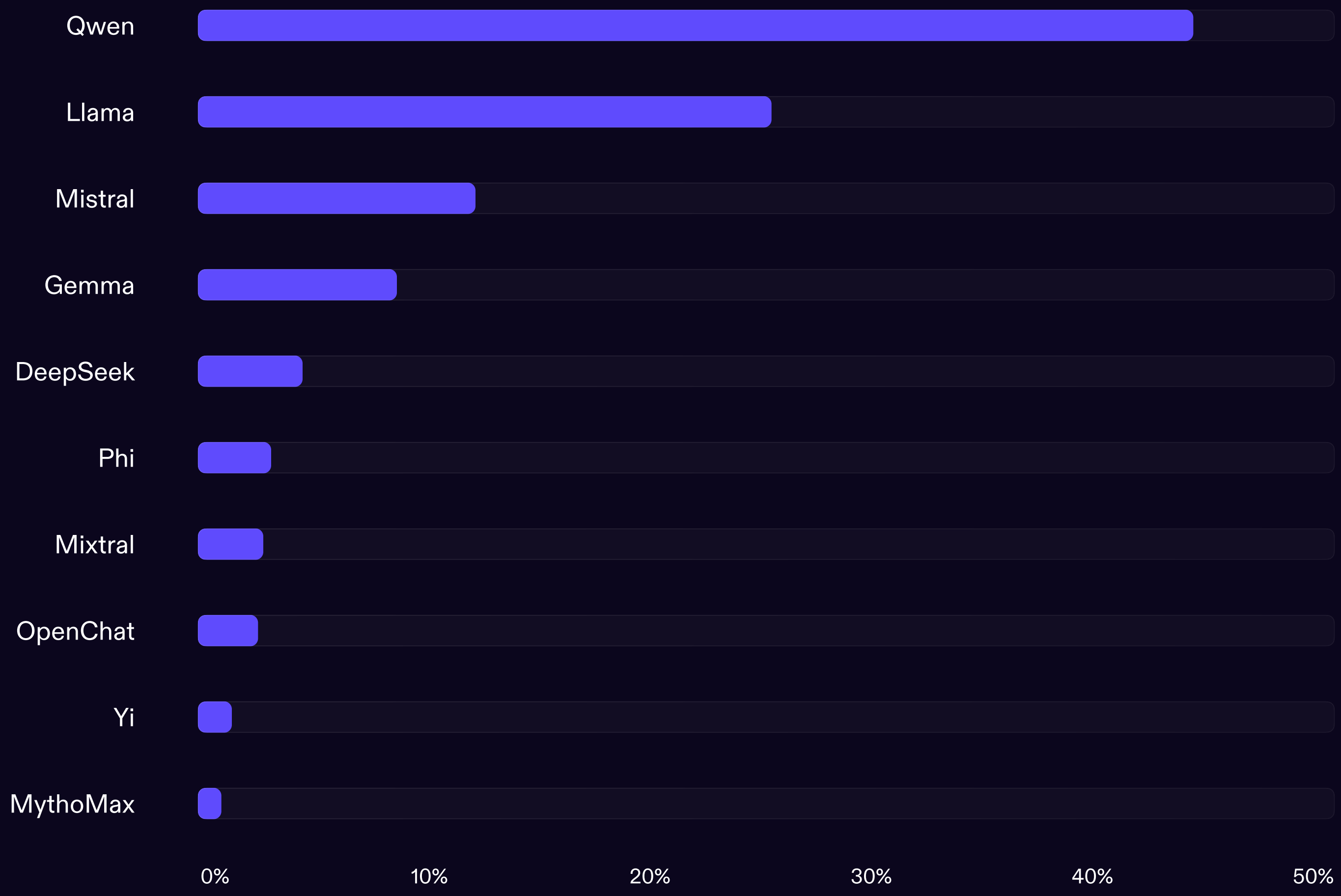
Llama's dominance may be shifting more permanently. Community adoption of Llama 4 has been slower than previous versions, and much of Llama's current popularity stems from extensive community fine-tuning of Llama 3. Newer architectures may continue to erode this position. We're seeing this shift in our data: Llama 4 (the most recent version) is used by only a handful of endpoints, while the vast majority of Llama usage is still on version 3.x. Even Llama 2 has 3x the adoption of Llama 4.

3x

Llama 2 vs. 4

If you're choosing an open-source LLM for production workloads, there are now several strong contenders worth evaluating.

Top Open Source LLMs



Share of Open Source LLM Usage (by number of endpoints)

On the infrastructure side, vLLM has become the de facto standard for LLM serving, with half of text-only endpoints running vLLM variants. Its optimizations for throughput and latency have made it the obvious choice for self-hosted inference, and its adoption suggests that practitioners have largely moved past the experimentation phase for serving infrastructure.

Chatbots remain one of the most visible applications of large language models, with businesses deploying them for customer support, internal knowledge bases, and interactive assistants. The rise of open-weight models like Llama, Qwen, and Gemma has made it increasingly practical to self-host conversational AI rather than relying solely on API providers.

Code generation is another emerging segment of text workloads. Users are building self-hosted coding assistants, alternatives to tools like GitHub Copilot that offer greater privacy and cost control. Qwen's Coder models dominate on Runpod, with the 30B and 32B parameter variants being the most popular.

Self-hosting gives you control over costs, latency, and data privacy, and with Runpod's serverless infrastructure, you can scale chat endpoints up or down based on actual demand.

1/2

of text-only endpoints running vLLM variants

What GPUs are Developers Choosing?

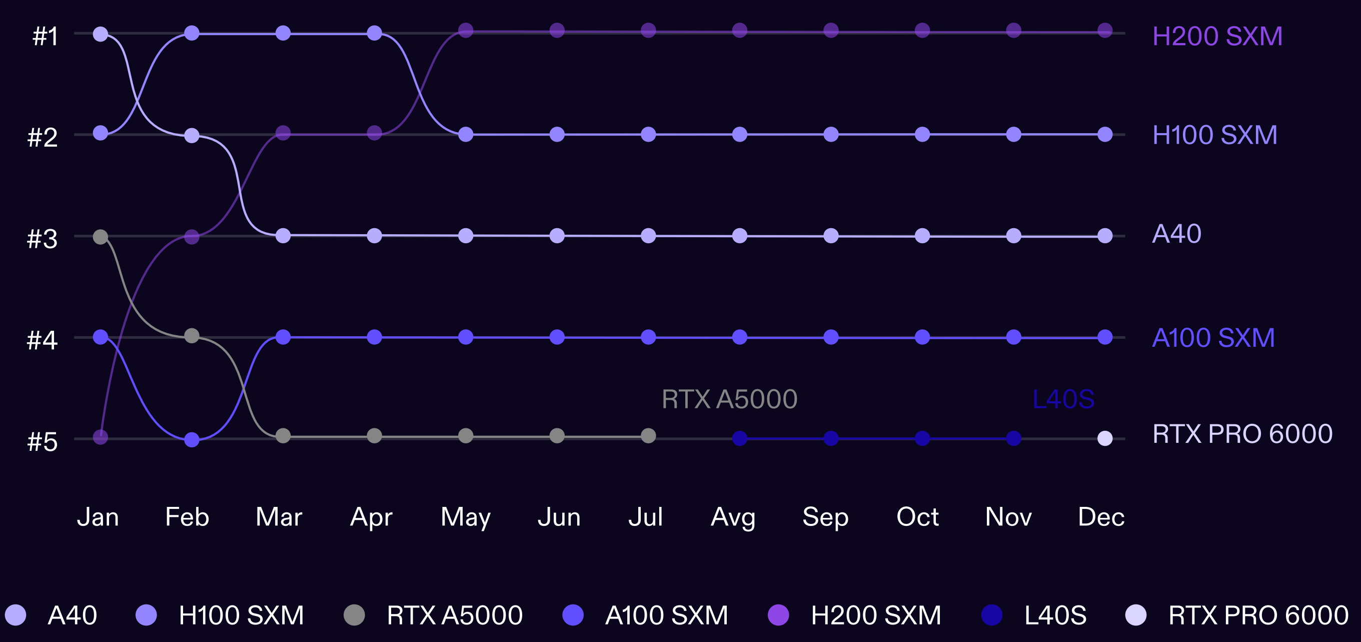
> The H100 continues to be a cornerstone of Runpod's platform, with usage growing by 68% throughout the year.

Its widespread adoption reflects both its proven performance and the maturity of the software ecosystem built around it. For many production workloads, the H100 remains the reliable choice.

But the next generation is arriving fast. The H200 SXM overtook the H100 SXM in May in terms of GPU hours and has held that position since, signaling that the Hopper to Hopper+ transition is well underway. For memory-intensive workloads like large language models and long-context inference, the H200's increased HBM capacity offers meaningful performance gains that justify the upgrade.

+68%
H100 usage

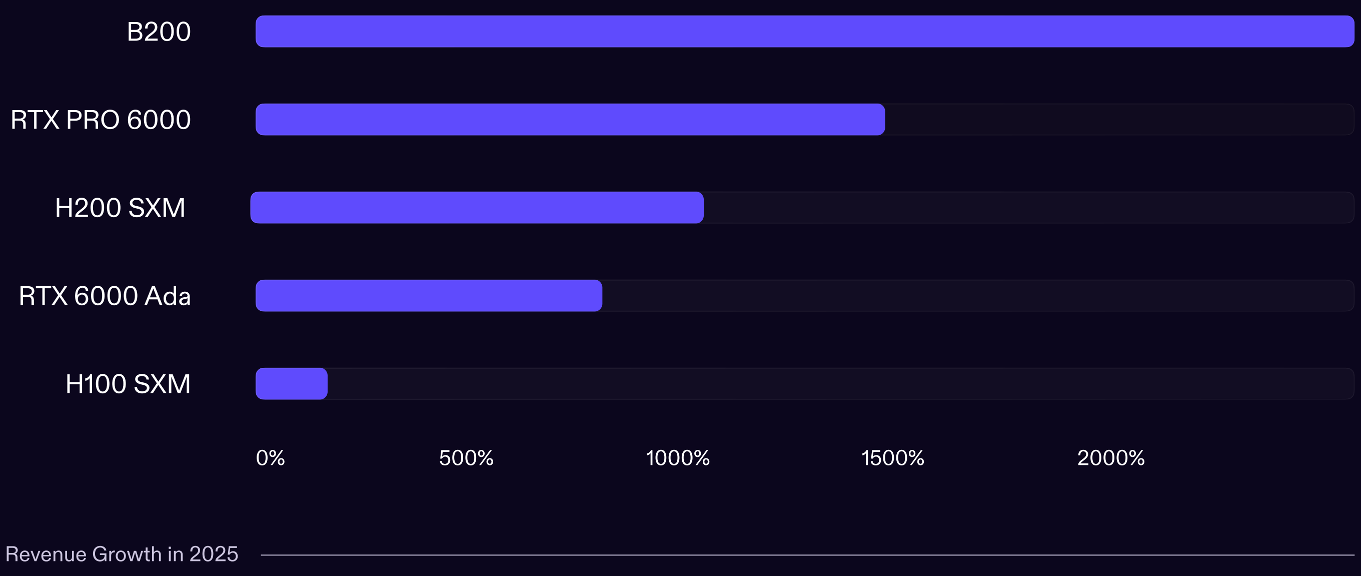
Top GPUs in 2025



Ranking (by Price-Weighted GPU Hours)

On the Blackwell front, Runpod invested heavily in systems software and infrastructure engineering in 2025, working closely with NVIDIA to optimize Blackwell performance, and that investment is paying dividends. The B200 emerged in April, and usage scaled by 25x by the end of the year. Early adopters are already seeing benefits for training and large-scale inference, and as Blackwell availability expands, this momentum is set to accelerate.

Fastest Growing GPUs in 2025



For users needing professional-grade graphics performance (particularly for image and video generation pipelines) the RTX PRO 6000 is a popular option. And 24GB cards remain a powerhouse for ComfyUI workflows, smaller model inference, and rapid prototyping, making them an accessible entry point for practitioners who don't need the memory footprint of larger GPUs.

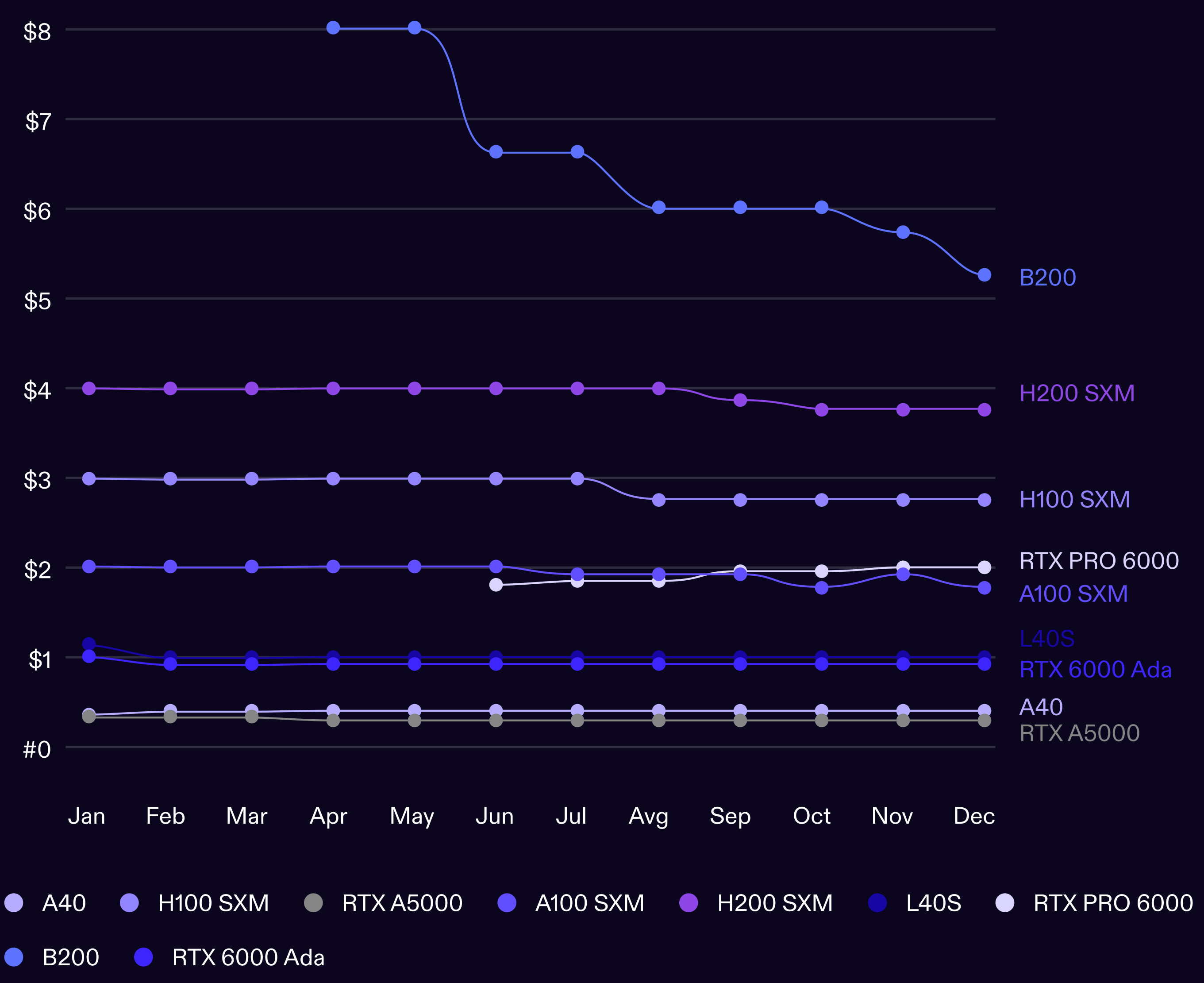
The right GPU choice depends on the specific workload: H100s for battle-tested compatibility, H200s for memory-bound inference, B200s for cutting-edge performance, and RTX PRO 6000s for cost-effective media generation.

GPU pricing trended downward across the board in 2025, reflecting growing access to supply. The steepest drop came from the B200, which fell 35% from $7.99 to $5.19 as production scaled up. The H200 SXM and H100 SXM each came down 10%, while mid-tier options like the L40S and A40 held largely flat.

-35%

B200 price

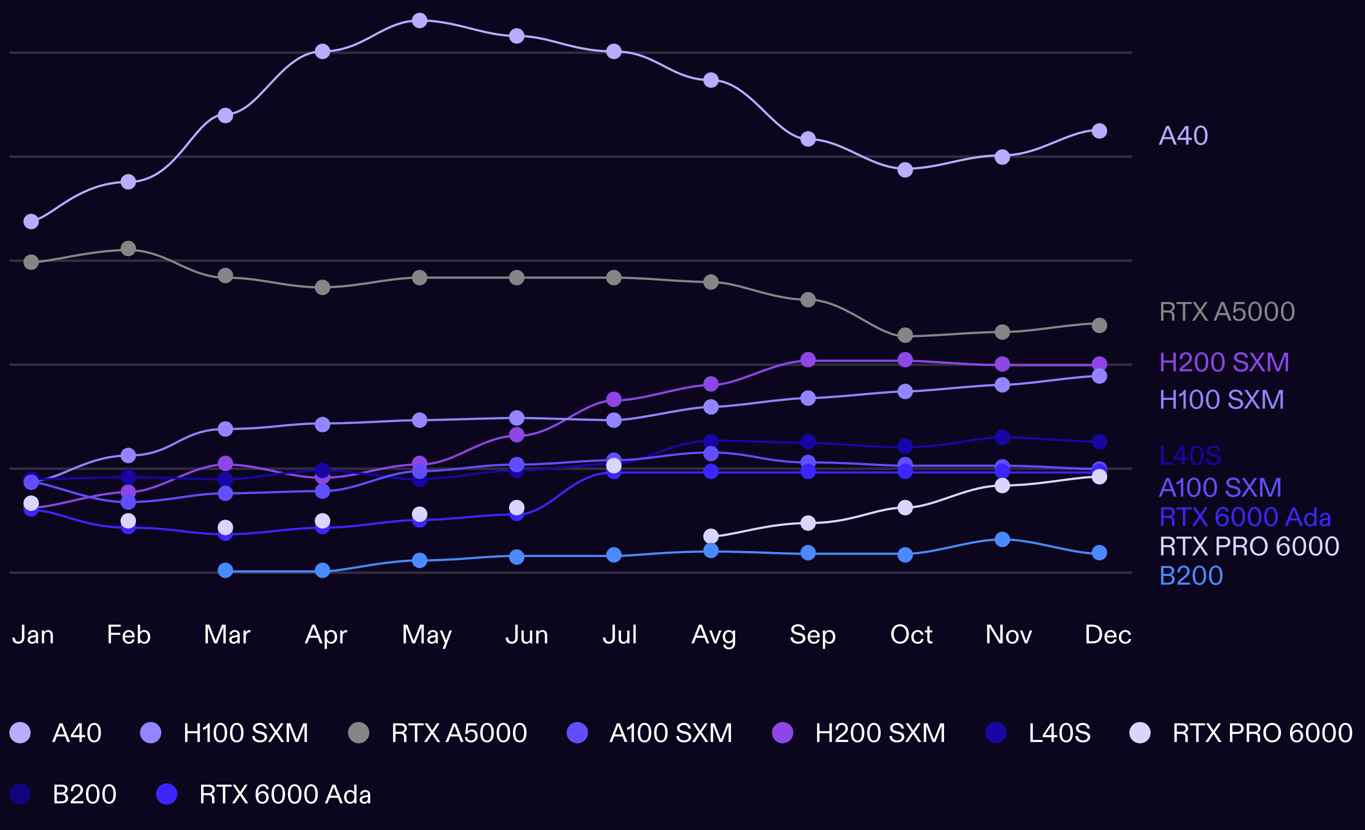
GPU Pricing Trends in 2025



Hourly GPU Price

On the supply side, the A40 remains the most widely available GPU on the platform, and its supply grew over 60% over the year. The most notable expansion came from NVIDIA's Hopper family: H100 SXM supply more than tripled, reflecting Runpod's continued investment in high-end data center hardware.

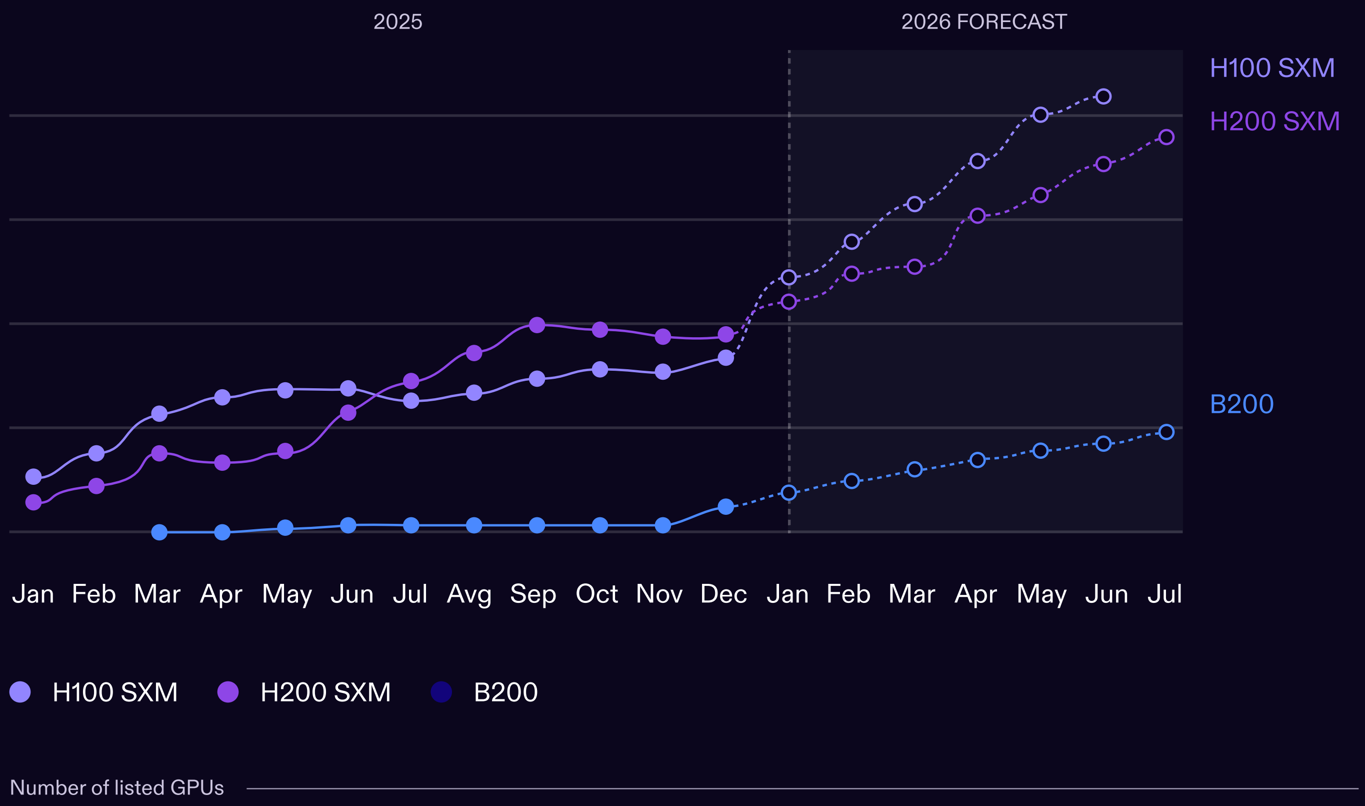
Supply in 2025



Number of listed GPUs

Looking ahead, H100 SXM and H200 SXM supply are both projected to roughly double by mid-2026 as Runpod continues scaling its Hopper infrastructure. B200 supply is expected to nearly quadruple as Blackwell adoption accelerates.

Forecasted Supply in 2026



Emerging tools and technologies

> Agentic AI has been one of the most talked-about trends in the industry this year, with frameworks like LangChain, LlamaIndex, CrewAI, AutoGen, Semantic Kernel and Mastra enabling developers to build autonomous, multi-step AI workflows.

We've started to see these frameworks appear among our serverless users, alongside a broader base of LLM inference infrastructure: vLLM deployments, function-calling models, and chat completion endpoints, all of which can serve as the backbone for agentic applications. The space is still early, but the building blocks are in place.

RAG has emerged as the practical approach to grounding LLM responses in domain-specific knowledge. Among our serverless users building RAG pipelines, we see embedding models like BGE and Qwen Embeddings used to vectorize documents, paired with vector stores such as FAISS and Chroma for efficient similarity search.

RAG is the starting point for retrieval workflows

Orchestration frameworks like LangChain and LlamaIndex help tie these components together into end-to-end retrieval workflows. For teams looking to make their LLMs actually useful on proprietary data, RAG has become the starting point.

Parameter-efficient fine-tuning methods like LoRA and QLoRA have transformed how developers customize models in the last few years, enabling meaningful adaptation without the compute requirements of full fine-tuning. This shift has made model customization accessible to a far broader range of users, and we see that reflected in our serverless data: LoRA appears commonly in fine-tuning endpoints, often paired with tools like Unsloth for optimized training and the PEFT library for streamlined workflows. If you've been hesitant to fine-tune because of cost or complexity, these techniques have fundamentally changed the calculus.

LoRA is commonly used by fine-tuning endpoints

Conclusion: The Infrastructure Layer for What's Next

The patterns in our data point to an AI ecosystem that's maturing rapidly. Model choices are consolidating around clear winners: Stable Diffusion for images, Wan for video, Whisper for transcription, Qwen for self-hosted LLMs. Tooling is standardizing on ComfyUI for visual workflows and vLLM for text inference. And practitioners are moving beyond simple generation toward sophisticated multi-model pipelines that combine image, video, audio, and text capabilities.

At the same time, the market remains dynamic. Flux is challenging Stable Diffusion's dominance. The H200 and B200 are reshaping GPU infrastructure. Text-to-speech is wide open for disruption. And agentic AI is just getting started.

What's clear is that building production AI requires flexible, scalable infrastructure that can keep pace with a fast-moving field.

The next wave of AI is being built right now.

Get started on Runpod >