



State of AI Compute

What 1.2 million developers run
when compute gets scarce

FALL 2026

Table of Contents

02	Executive summary
03	Key findings
04	Methodology
05	Part 1: The new economics of AI compute
08	Part 2: Compression, customization, and right-sized compute
11	Part 3: Open weights and frontier APIs often run together
15	Part 4: What developers actually run
17	Part 5: Agents are becoming durable infrastructure workloads
19	Part 6: AI compute is broadening beyond AI companies and the U.S.
22	Part 7: How Runpod is meeting demand
25	Part 8: Looking back at the last supply calls
26	Conclusion



Executive summary

Teams have more ways to shape an AI workload than they did a year ago: more control over what runs, at what size, on what hardware, and at what cost. Hardware scarcity drove much of that change. High-bandwidth memory is committed into 2027, and some planned data center capacity is waiting on grid connections. Those constraints forced innovation, and the past year produced plenty of it. When the direct path is hard to source, teams build around it.

The flip side is that optimization is no longer optional. Precision, model size, and hardware class now determine what a workload costs to run, how fast it responds, and how good the output is. Teams must make those choices deliberately, validate them against their quality bar, and maintain them over time. The same expansion in choice that lets a team tune a workload precisely also makes tuning part of the job. This report looks at how those decisions are being made.

There is very little public data on what teams actually deploy in production. Runpod sees it directly. Developers choose the model, the precision, and the hardware, workload by workload. The platform is a record of those decisions. This second edition of the *State of AI Compute* draws on anonymized Serverless traffic, Pod activity, GPU consumption, revenue, and sign-up trends across 1.2 million accounts worldwide.



Key findings

01

Agent-provisioned resources generate disproportionate revenue

Agents account for 8.4% of Runpod users, but resources they create account for 24% of revenue, or roughly 2.9x the platform average per user. Their revenue share rose from 10% to 12% to 24% across the three monthly snapshots.

02

Workflow choices differ sharply by modality

Among video Pods, 62.9% only process footage that already exists, and 85.2% do post-production of some kind. Image workloads are the opposite, with generation in 82.2% of Pods, while audio is primarily transcription.

03

The fastest growth is coming from outside the AI industry

Eight of Runpod's ten fastest-growing industries are not AI-specific, although AI companies remain the heaviest users.

04

Frontier APIs and local models often coexist

Among Pods that reference a frontier model, 80% also run open weights or a local serving stack.

05

Software efficiency is changing hardware requirements

Quantization scales with model size: 11.6% of Pods serving models under 8 billion parameters use a quantized build, rising to 32.5% between 8 billion and 70 billion, and 40.7% above 70 billion. Teams compress large models to fit available memory; small models are more often customized instead.





How quantization works

Quantization is a way of representing model weights with fewer bits. That can lower memory requirements and make a model fit on smaller hardware, but teams still need to validate quality for their use case.

Methodology

This report draws on anonymized Runpod platform data from 1.2 million users collected from April 1, 2025 through Aug. 31, 2026, supplemented by internal fleet forecasts where noted. The unit of analysis varies by section and may include users, new business sign-ups, revenue, Pods, endpoints, or requests. Each finding should be read using the denominator and time period stated alongside it.

The findings describe activity observed on Runpod. They are not representative of every AI workload running across hyperscalers, private enterprise clusters, AI labs, or sovereign infrastructure.



The new economics of AI compute

The economics of building with AI go beyond model selection. The harness and serving stack shape how requests are batched, cached, routed, and validated. GPU memory, availability, storage, networking, and power determine what can run and what it costs.

At a glance

2027

SK Hynix expects the memory shortage to reach its most severe point

0-15%

Runpod price adjustments from June to August 2026

+7% vs.
+31%

Change in Runpod's A100 SXM and the newer B200 list prices, respectively, from January to August 2026

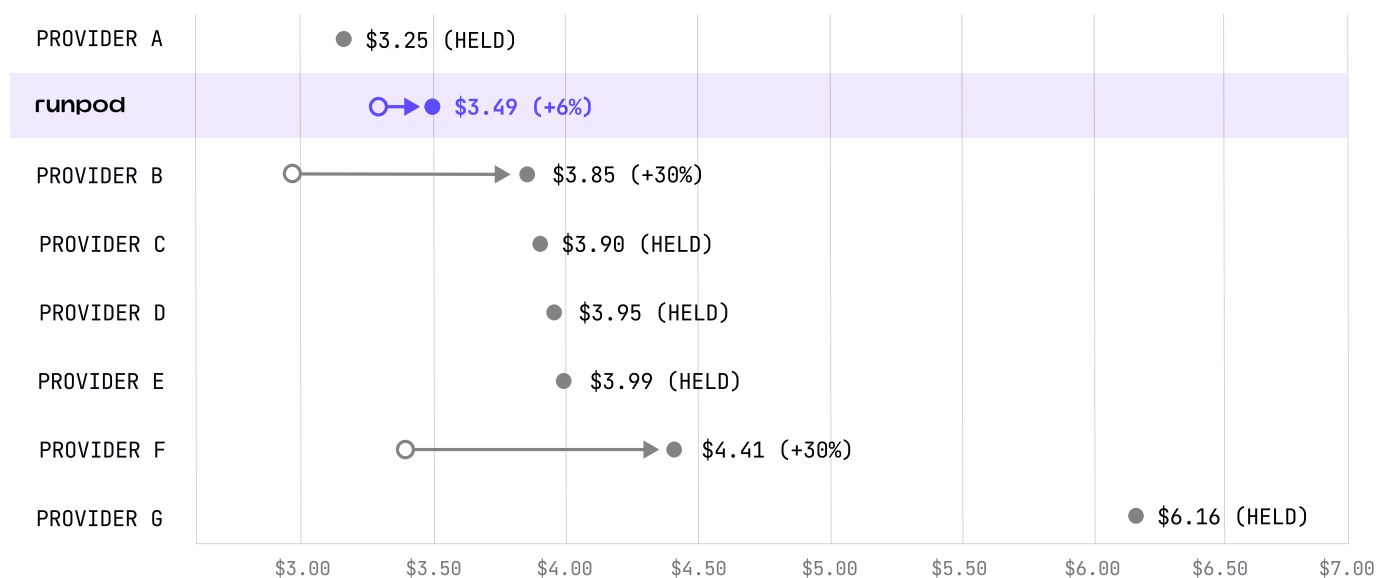
Pressure is building across several parts of the hardware supply chain. High-bandwidth memory is effectively sold out into 2027, and SK Hynix expects the broader memory shortage to peak in 2027 and persist toward 2030. Capacity for CoWoS, the advanced packaging technology used to integrate GPUs with HBM, is also booked well in advance. Some planned datacenter capacity remains stalled by grid constraints.

The pressure extends beyond flagship GPUs. DDR5 server memory remains in short supply as memory manufacturers prioritize higher-margin HBM. Enterprise flash prices have also risen as buyers secure capacity. Teams building or expanding their own clusters must account for procurement time, upfront capital, and component-price volatility.



A Runpod review of published list prices found that flagship GPU instances became more expensive across several providers between June and August 2026, although changes varied by GPU, region, and contract structure. Runpod's adjustments over the same period ranged from 0% to 15%. Across the longer January-to-August window, Runpod's A100 SXM list price rose 7%, compared with 31% for the newer B200. The largest increases affected flagship GPUs built around constrained components such as HBM; midrange and older data-center GPUs moved less.

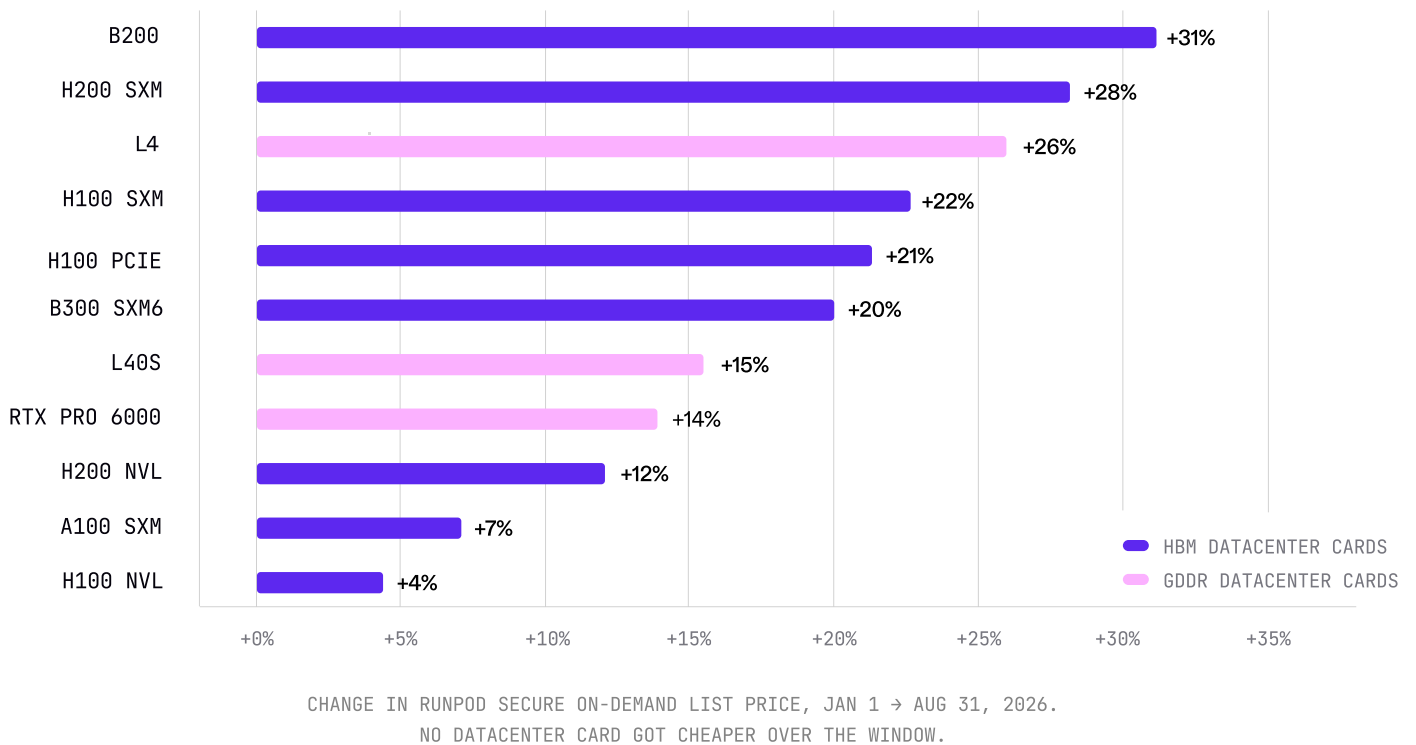
THE MARKET IS REPRICING UPWARD



Lower-priced models are also handling more inference traffic. On OpenRouter, DeepSeek's share of combined input and output token volume rose from 9% in January to 18% in June 2026. DeepSeek V4 Flash was available for \$0.09 per million input tokens and \$0.18 per million output tokens through its least expensive route. The [OpenRouter analysis](#) measures token volume, not request share, spend, or cost per successful task. It also does not establish that price caused the increase.



2026 LIST PRICE MOVES: THE HBM FLAGSHIPS LEAD THE REPRICING



Runpod data shows developers changing workloads as well as hardware. Teams are compressing models, selecting smaller models for specialized tasks, fine-tuning for narrower use cases, and matching workloads to GPU memory classes. These approaches do not remove hardware constraints, but they can change a workload's memory, throughput, and utilization profile. The effect on quality and total cost depends on the deployment.

➔ Developer takeaway

Start capacity planning with the workload, not the GPU name. Define the required memory, architecture, latency, storage pattern, and tolerance for idle capacity before selecting hardware.



Compression, customization, and right-sized compute

Software can reduce the compute required for a useful result in several ways. Quantization changes how a model is represented; smaller models narrow the job; decision models remove free-form generation when an application only needs a defined answer.

At a glance

11.6% to
40.7%

Share of Pods serving quantized builds under 8 billion parameters versus above 70 billion

95.9% to
85.8%

Share of Qwen3 endpoints using 16-bit builds, while 8-bit rose from 2.0% to 10.9% and 4-bit from 3.1% to 8.8%

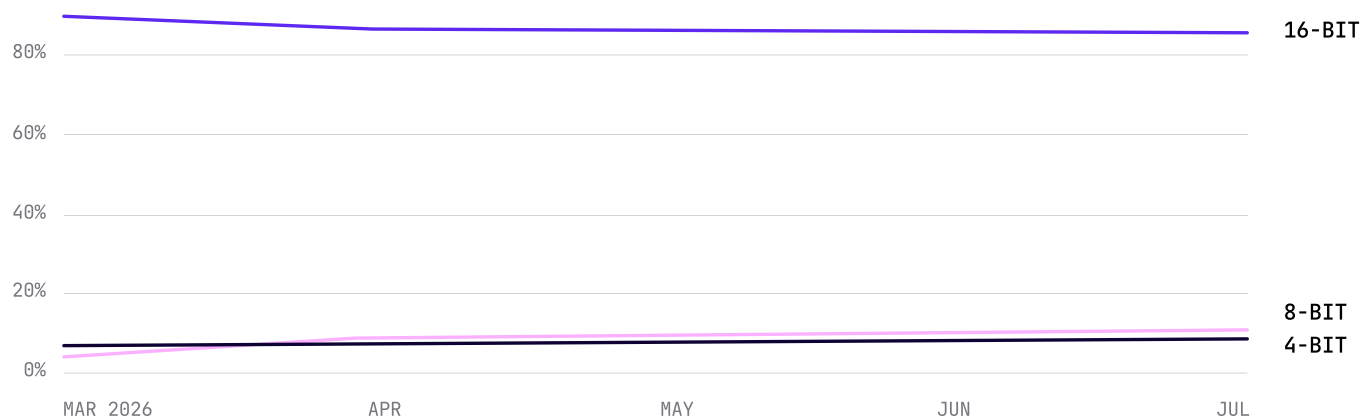
7.5% vs.
2.8%

Share of Pods used for training or fine-tuning among those naming only small models versus no models

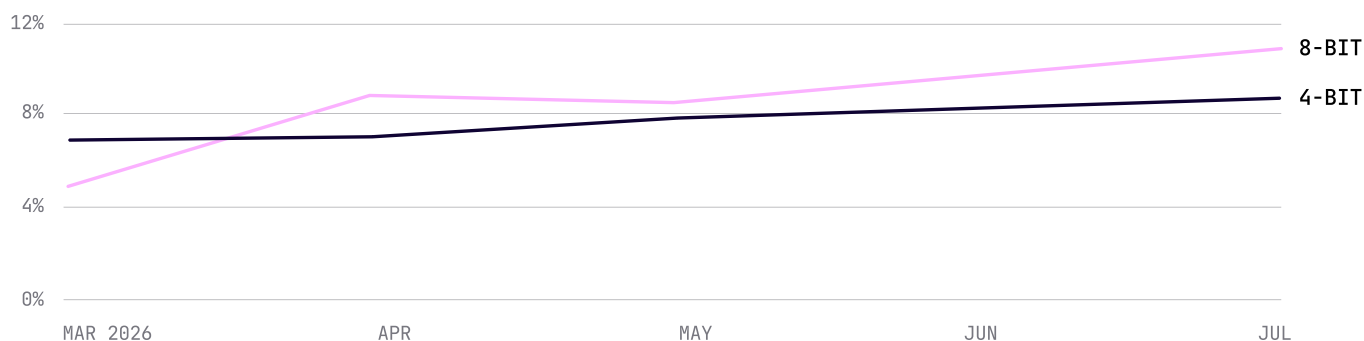
Among Pods serving models under 8 billion parameters, 11.6% use a quantized build. That rises to 32.5% for models between 8 billion and 70 billion and 40.7% for models above 70 billion. Quantization is more common among larger models, where memory requirements are higher.



PRECISION MIX AMONG QWEN3 ENDPOINTS



LABELED LOW-PRECISION BUCKETS, RESCALED



Small models address a different part of the efficiency problem. On Runpod, they perform narrow tasks within larger pipelines, including captioning, diarization, deduplication, redaction, product tagging, and validation. Examples include NVIDIA Parakeet, Florence-2, all-MiniLM, SmolVLM, and Moondream.

Small models also appear more often in environments that include training or fine-tuning. Among pods that name only small models, 7.5% include training or fine-tuning, compared with 2.8% of pods that name no small models. This is an association, not evidence that small models cause more fine-tuning. It suggests that teams frequently adapt them for specific tasks.



Post-period example

Jev narrows the task

Jev illustrates another efficiency strategy: using defined outputs instead of free-form generation when the task permits. TypeSafe AI released Jev on September 15, 2026. It returns developer-defined typed answers with probabilities and, at launch, cost \$0.042 per million input tokens with no output-token charge.

Some developers are also optimizing for latency. Speculative decoding and multi-token prediction appear in serving stacks across Qwen, DeepSeek, and Gemma. Some quantized builds retain multi-token prediction heads, combining lower-precision weights with a technique designed to increase decoding speed.

→ Developer takeaway

Use quantization when memory limits deployment and the lower-precision build meets quality requirements. Small models fit narrow, repeated tasks with clear evaluation criteria. Speculative decoding and multi-token prediction address latency and throughput.



Open weights and frontier APIs often run together

Frontier APIs and local models often appear in the same Runpod environment. Among Pods that name a frontier model, four in five also run local weights.

At a glance

4 in 5

Pods naming a frontier model that also run local weights

74.2%

Share of text endpoints running a Qwen model

1.7% to 25.4%

Growth in LTX's share of video endpoints over the report period

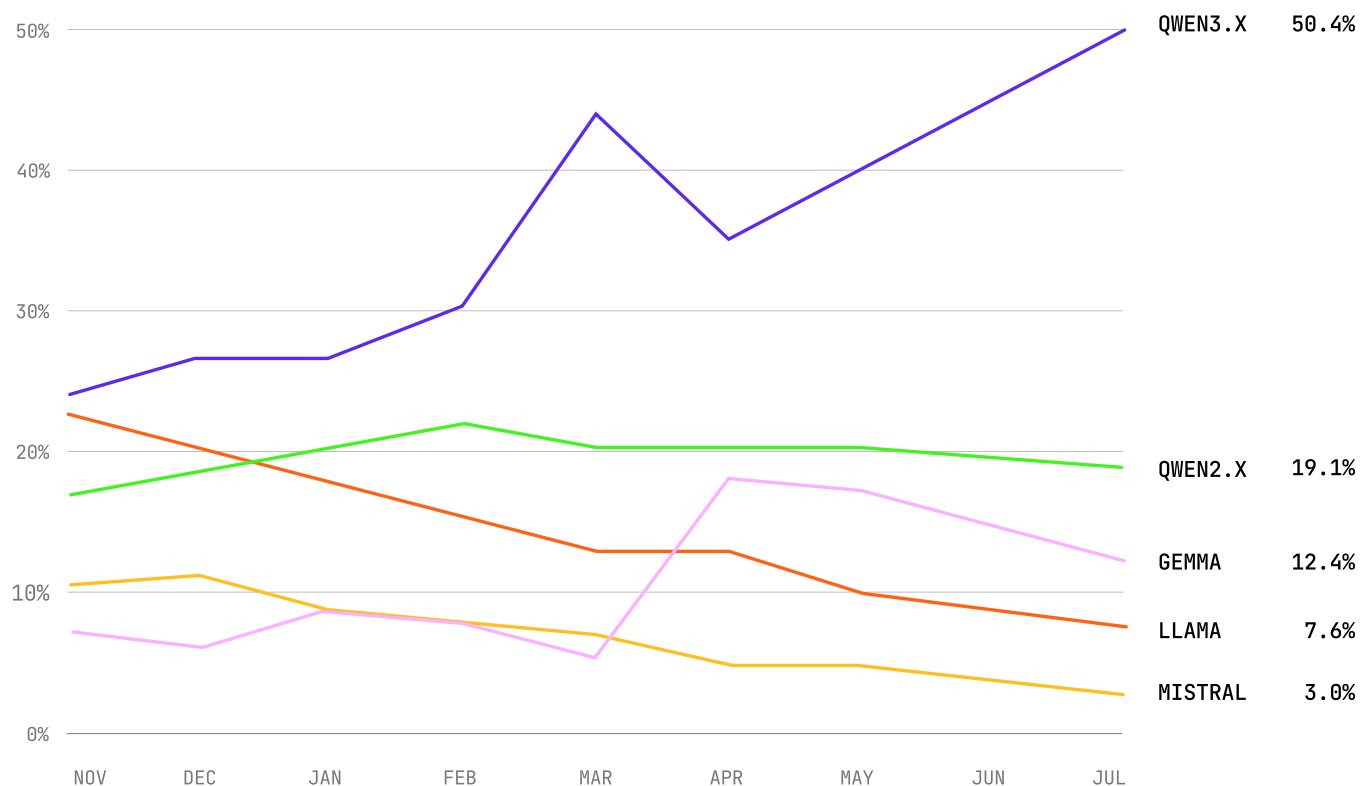
The data shows that some teams use frontier APIs and local models together. It does not show how traffic is divided between them or that one model type is replacing the other.

Among open-weight text models, Qwen has the largest presence on Runpod: 74.2% of text endpoints across versions. Qwen3.x rose from 24.3% to 50.4% since November 2025, while Qwen2.x reached 19.1%. Over the same period, Llama fell from 23.0% to 7.6% and Mistral from 10.8% to 3.0%. Gemma rose from 5.5% in March to 18.2% after the release of Gemma 4.



USAGE SHARE AMONG TOP TEXT MODELS

SHARE OF TEXT SERVERLESS ENDPOINTS USING MODEL FAMILY, NOV 2025 - JULY 2026



Some of these changes coincided with new model releases. From February to March, when Qwen3.5 open weights became available, Qwen's share of text endpoints rose from 61.0% to 69.2%. From April to May, when Qwen3.6 weights became available, its share rose from 69.2% to 72.1%.

Kimi shows how quickly the mix can change within one model family. Kimi joined Runpod Public Endpoints in late July 2026 and reached 14% of requests there. Among self-hosted Kimi Pods, K2.7 held 45% in July; by August, K3 held 60%. A new version took majority share of Kimi workloads in roughly a month. One common concern about running your own weights is that upgrades lag, and these deployments did not. This measures version share among Kimi workloads on Runpod, not Kimi's adoption across the broader market.

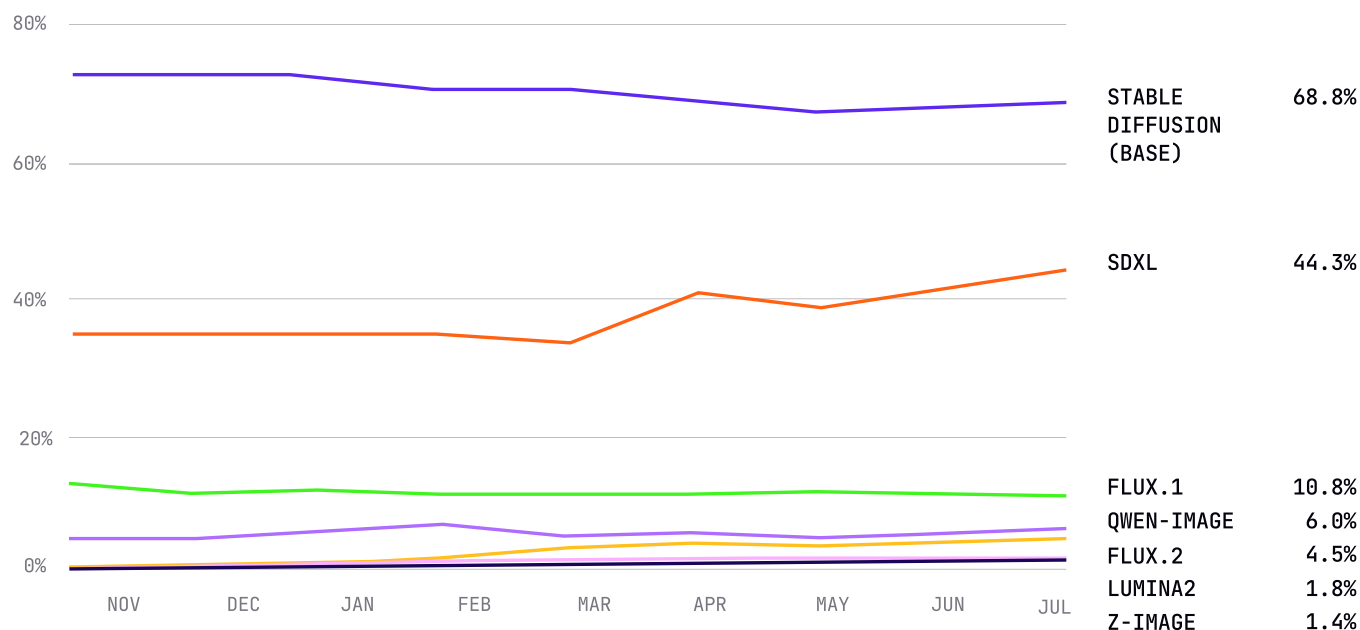
Model shares differ by modality. Base Stable Diffusion remains the most common image model, though its share fell from 73.3% to 68.8%. SDXL rose from 34.4% to 44.3%. Flux.2, Lumina2, and Z-Image are all smaller.

In video, Wan leads at 72.3%, down from 86.6%. LTX rose from 1.7% to 25.4%.

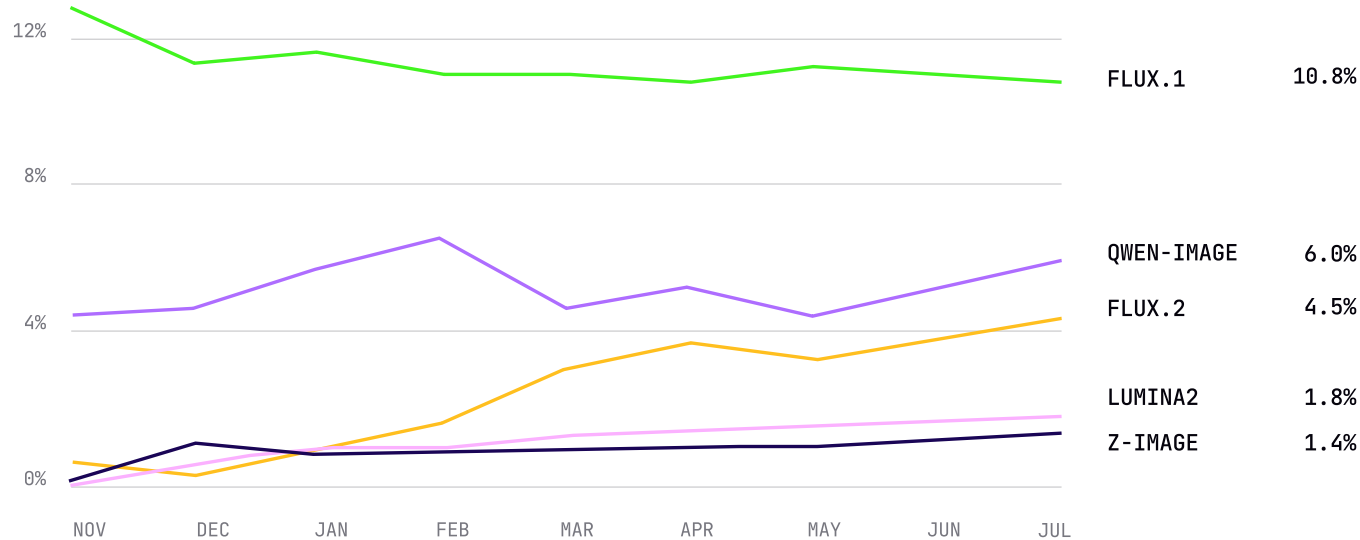


USAGE SHARE AMONG TOP IMAGE MODELS

SHARE OF IMAGE SERVERLESS ENDPOINTS USING MODEL FAMILY, NOV 2025 - JUL 2026



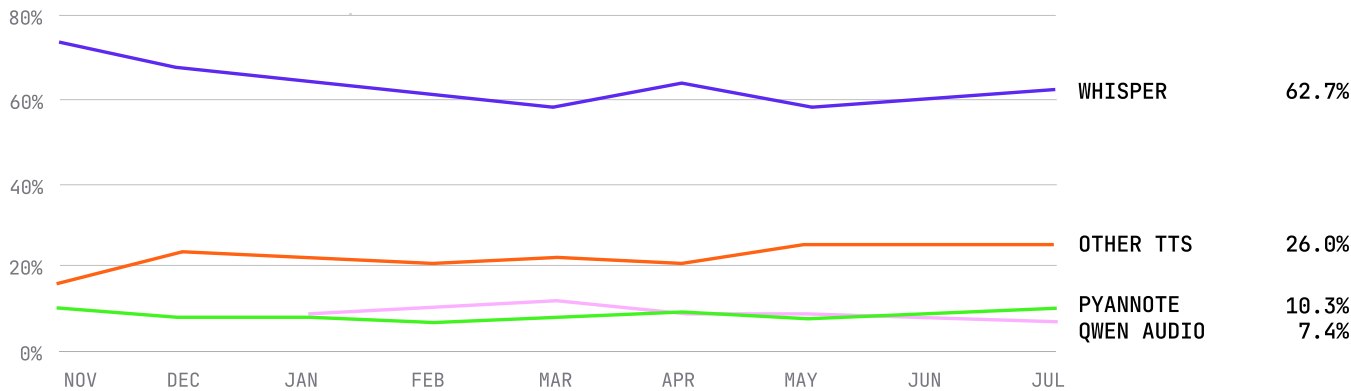
TOP NON-STABLE DIFFUSION IMAGE MODELS, RESCALED



Whisper leads audio at 62.7%, down from 74.0%. Text-to-speech models collectively rose from 16.7% to 26.0%, with no single TTS model accounting for most of that usage.

USAGE SHARE AMONG TOP AUDIO MODELS

SHARE OF AUDIO SERVERLESS ENDPOINTS USING MODEL FAMILY, NOV 2025 - JULY 2026



Several of the most-used open models come from Chinese labs. Qwen runs on 74.2% of text endpoints, Wan leads video at 72.3%, and Kimi reached 14% of Public Endpoint requests shortly after becoming available. Llama, which led open-weight text usage two years ago, is at 7.6%. The pattern is not uniform across modalities: Stable Diffusion and SDXL lead image workloads, and Whisper leads audio at 62.7%.

External platform data shows a similar rise in Chinese open models. OpenRouter reported that Chinese models surpassed U.S. models in token share in early June 2026. In March 2026, Hugging Face reported more than 113,000 Qwen derivatives. These measures are not directly comparable with Runpod’s endpoint share: OpenRouter measures hosted token volume, Hugging Face measures model derivatives, and Runpod measures models deployed on its infrastructure.

Cost, control, customizability

For business-critical workloads, open models are not just about ideology. Teams may want models they can run inside their own infrastructure, tune for a specific task, keep available even when external providers change policies, and govern around their own data requirements.



PART 4

What developers actually run

“Generative AI” is not one workload. On Runpod, video, image, and audio workloads look very different.

At a glance

85.2%

Share of video
Pods that perform
post-production

82.2%

Share of image
Pods that perform
generation

43.6%

Share of audio
Pods that perform
speech-to-text

Video is mostly post-production. Among video Pods, 62.9% work on footage that already exists and never generate new footage. Another 22.3% both generate and post-process. Only 11.2% generate video and do no post-production. In total, 85.2% of video Pods touch post-production in some form.

Upscaling is the most common video workflow at 77.0%, followed by enhancement at 24.7% and frame interpolation at 10.3%. ComfyUI appears in 74.3% of upscaling work. Upscaling and frame interpolation can require substantial memory and throughput, particularly when processing longer assets frame by frame.



Image workloads are more generation-heavy. Generation appears in 82.2% of image Pods, and 61.8% generate without post-processing. Image analysis, including captioning, image-to-text, OCR, and detection, is the second-largest workflow at 18.1%. Upscaling appears in 14.7% of image Pods, enhancement in 7.3%, and face swapping and restoration together in 3.3%.

Audio follows a different pattern. Speech-to-text is the largest audio workflow at 43.6% of audio Pods, ahead of text-to-speech at 28.4%. Audio processing is 26.7%, and diarization is 14.2%.



Developer takeaway

Plan infrastructure around the workflow, not the “generative AI” label. A video upscaler, image generator, OCR pipeline, transcription system, and diarization workflow have different requirements for memory, latency, storage, and idle capacity.



Agents are becoming durable infrastructure workloads

Most Pods running agent frameworks on Runpod are short-lived, but they last longer than the platform baseline. Agents also run hybrid model stacks, and they are beginning to create infrastructure resources directly.

At a glance

63%

Agent-framework Pods that use external frontier APIs while hosting open models locally

3.7x

Greater likelihood that an agent Pod runs longer than one month

24%

Runpod revenue attached to resources created by agents

Agent frameworks observed on Runpod include LangChain, LangGraph, LlamaIndex, CrewAI, n8n, and OpenHands. Most are retrieval and orchestration systems: knowledge-base assistants, customer support workflows, and document question-answering over vector stores. Others include a multilingual voice collections agent, a health care documentation workflow, and a GitHub pull request reviewer.

Among Pods running agent frameworks, 63% orchestrate against external frontier APIs while also hosting open models locally. In these workloads, Runpod hosts part of the agentic application, not only a model.



Most agent Pods are still short-lived, as are most workloads on the platform. Compared with the platform baseline, however, agent Pods are 3.7 times more likely to run longer than a month: 7.7% versus 2.1%. They are 2.6 times more likely to run longer than a week, and less likely to disappear within the hour (17.5% versus 30%).

Runtime does not prove that a workload is in production, but it's one indication. Workloads that persist for weeks or months are less likely to be one-off tests.

Agents are also managing infrastructure resources. Resources created by agents account for 24% of Runpod revenue, up from 12% a month earlier and 10% the month before that. Agents are 8.4% of Runpod users, up from 4.2% and 2.4% over the same two months.

Precision note

Revenue from agent-created resources means that an agent, rather than a human, created the associated Pod or Serverless endpoint. The agent is not the customer, budget owner, or legal buyer.

Agent workloads on Runpod also include reinforcement learning: robotics and embodied simulation, trading agents, and language-model post-training using PPO, DPO, ORPO, RLHF, and reward-model fine-tuning. The tooling includes Gymnasium, Stable Baselines3, MuJoCo, Brax, Ray RLlib, and rsl_rl. These workloads are more specialized than retrieval and chat orchestration.



AI compute is broadening beyond AI companies and the U.S.

AI companies remain Runpod's heaviest users, but the fastest growth is coming from other industries and markets.

At a glance

61.6%

Share of revenue from
AI companies

8 of 10

Fastest-growing
industries that are not
AI-specific

186

Countries
represented across
Runpod's user base

AI companies account for 30.2% of Runpod users and 61.6% of revenue. They remain the center of gravity for spending, but they are only 25.9% of new business sign-ups.

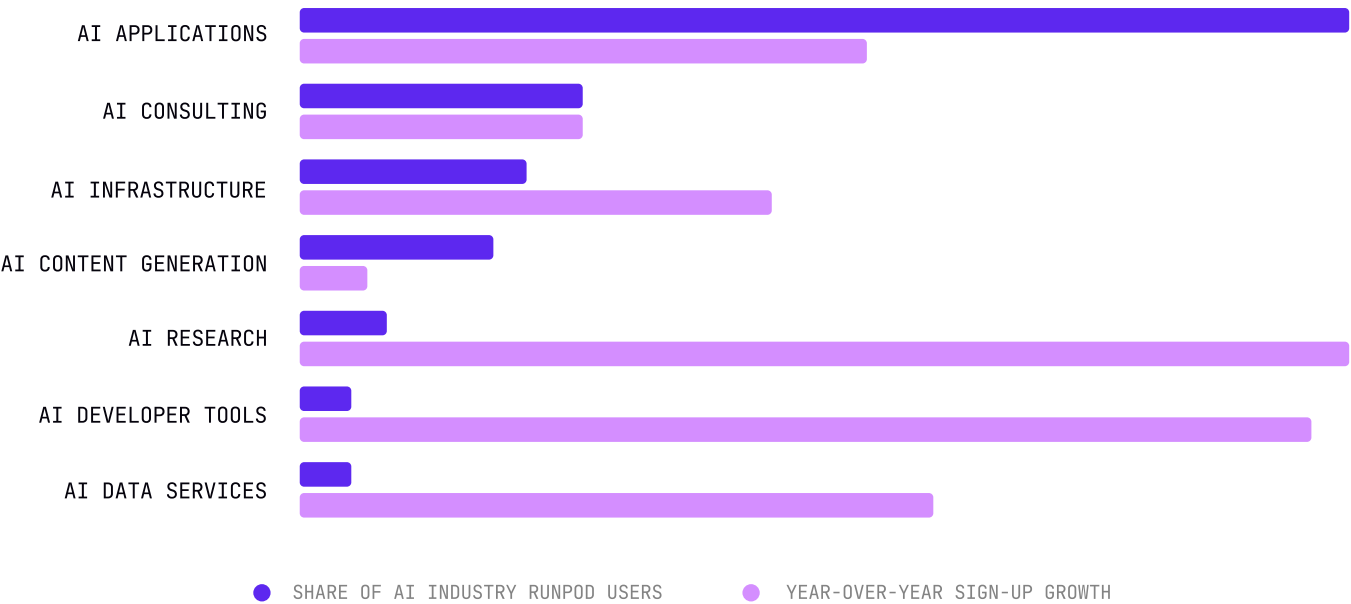
Government, fintech, and manufacturing are the three fastest-growing industries on Runpod. Eight of the ten fastest-growing are not AI-specific, and none of the top five are. AI companies are not becoming less important. Demand for accelerated compute is also growing among organizations whose core products and services are not AI.

Government growth skews to the U.S. but is not limited to it. Among government users added since February 2026, 36.2% are in the U.S., compared with 18.6% of Runpod users overall.



Within the AI category, applications are the largest subcategory at 53.3% of AI industry users, followed by consulting at 14.1%, infrastructure at 11.3%, and content generation at 9.7%. The fastest-growing subcategories are AI research, AI developer tools, and AI data services.

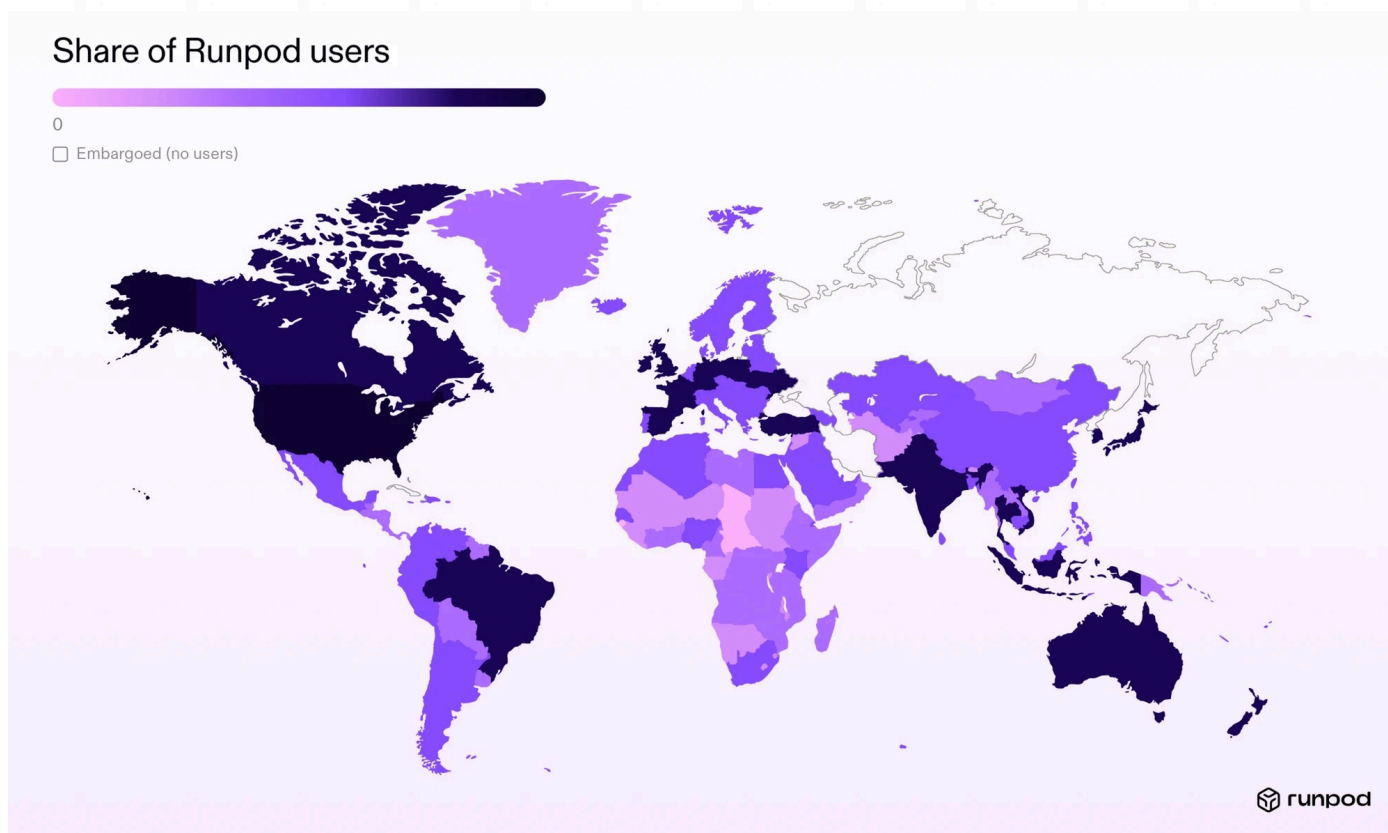
AI INDUSTRY SUBCATEGORIES



The user base is also geographically distributed. Runpod has 1.2 million users across 186 countries. The U.S. remains the largest market at 18.6% of users. India is the largest non-U.S. market at 9.1%, followed by Germany at 5.0%, Brazil at 3.6%, Korea at 3.3%, and the U.K. at 3.2%.

Asia is the fastest-growing continent on Runpod and includes four of its ten fastest-growing countries. Japan is the third-fastest-growing country and now Runpod’s seventh-largest user base. Indonesia ranks fourth by growth, Singapore sixth, and Hong Kong ninth. Southeast Asia accounts for 9.8% of all new sign-ups, while East Asia accounts for 9.5%.

Revenue remains more geographically concentrated than users. The U.S. accounts for 18.6% of users but 38.8% of revenue. New markets tend to add users before they add spend, so user share is the earlier signal.



→ Developer takeaway

The potential audience for AI models, developer tools, data services, and deployment infrastructure increasingly includes companies whose primary business is not AI.



How Runpod is meeting demand

Runpod is responding to supply constraints by expanding its GPU mix, dividing high-memory cards into smaller instances, and investing in the storage and networking required to support larger workloads.

At a glance

96GB →
4 × 24GB

One example
of partitioning
a larger GPU

36.9%

Share of fine-tuning
Pods using at least
80GB of VRAM

2.50x

Forecast expansion of
Runpod's H200 fleet
by early December

Runpod's GPU mix has shifted toward datacenter-grade hardware in 2026 as customers run larger models, longer training jobs, and more demanding inference pipelines.

The H100 SXM remains the backbone of many production workloads. The H200 SXM supports memory-bound inference and long-context workloads where additional HBM can change the feasible batch size or model class. Within Runpod's fleet, B200 adoption expanded during the first half of the year. B300 SXM6 is now live with 800 Gb/s InfiniBand networking between servers and more than 740 Gb/s of GPU-to-GPU bandwidth for multi-node training.



The RTX PRO 6000 Blackwell Server Edition supports media-generation and rendering pipelines that require graphics capability alongside ML compute. AMD MI300X is drawing interest as ROCm matures, while the A100 SXM continues to serve a price-performance segment for optimized inference and fine-tuning.

No single GPU class fits every workload. The appropriate option depends on memory requirements, architecture, software support, availability, and price.

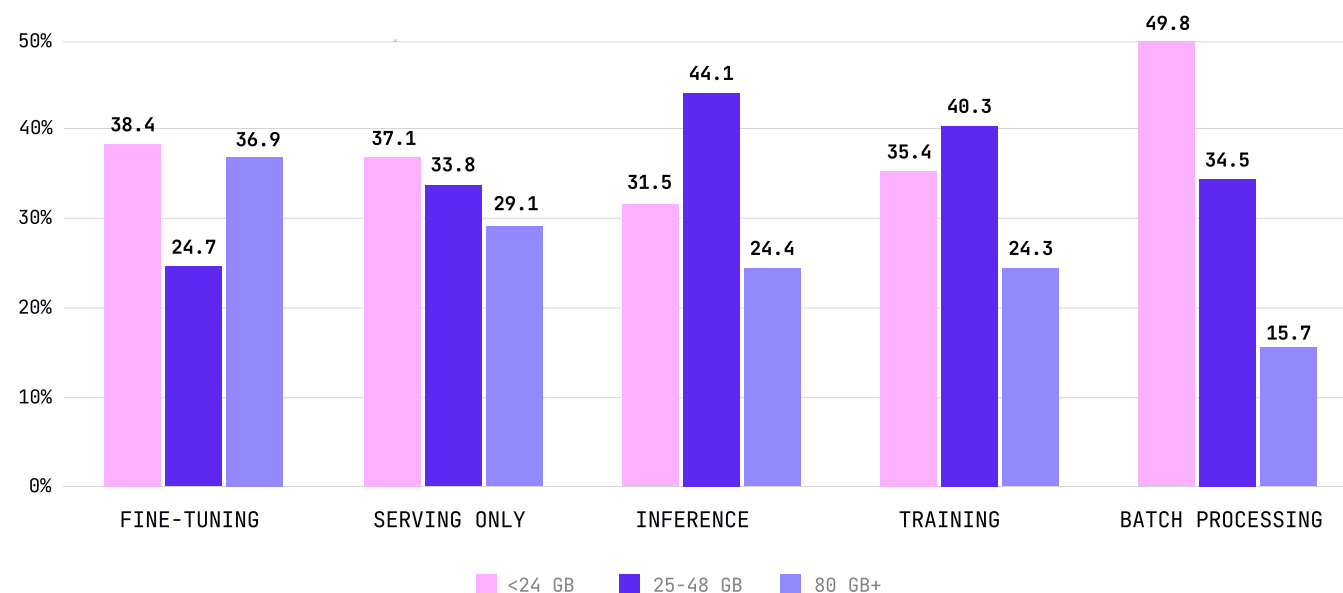
GPU partitioning is one way Runpod matches available supply to different memory requirements.

Multi-Instance GPU technology divides a large card into isolated, smaller GPU instances at the hardware level. A 96GB RTX PRO 6000 can provide four independent 24GB instances, and B200 and B300 capacity can be served through 24GB and 48GB Serverless endpoints. These partitions give a workload the newer architecture and its low-precision formats without reserving the entire card.

Partitioning complements the shift toward lower precision described in Part 2. Quantization can reduce a model's memory footprint, while partitioning opens smaller, isolated portions of newer GPUs. Runpod data shows lower-precision Qwen3 endpoints becoming more common over time, but it does not show that quantization is primarily moving those deployments into smaller VRAM classes.

VRAM selection also varies by workload. Fine-tuning is the workload most likely to use at least 80GB, accounting for 36.9% of fine-tuning Pods, mostly on A100, H100, and H200 GPUs. Inference most often uses the 25-48GB tier, while just under half of batch-processing workloads run on 24GB or less.

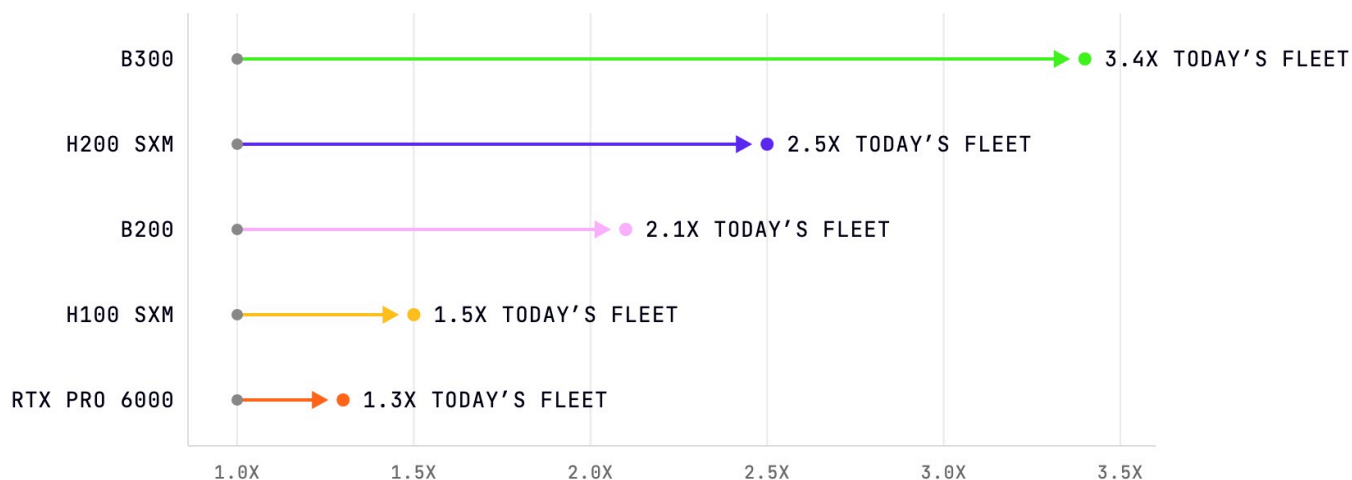
VRAM TIER BY WORKLOAD TYPE



Modality introduces another difference. Text has the highest share of Pods using at least 80GB, at 40.1%. Video and image workloads both have a median of 32GB, but video has a heavier upper range: 30.5% of video Pods use at least 80GB, compared with 19.7% of image Pods. The 75th percentile is 80GB for video and 48GB for image. This pattern is consistent with the memory demands of operations such as video upscaling and interpolation, but the data does not establish those operations as the cause.

Runpod's internal forecast calls for capacity expansion by early December: 2.5 times the current H200 fleet, 2.1 times B200, 1.5 times H100 SXM, and 1.3 times RTX PRO 6000. These are internal forecasts, not measurements of demand across the broader AI infrastructure market.

THREE FLEETS DOUBLE BY EARLY DECEMBER



GROWTH BY EARLY DECEMBER, AS A MULTIPLE OF TODAY'S FLEET

FLEET PLANS TRACK THE P50 DEMAND FORECAST, 3.5-MONTH HORIZON (DEC 7), AUGUST 2026. MULTIPLES OF CURRENTLY VERIFIED SUPPLY.

Additional supply is expected through Q3 and Q4, weighted toward newer cards such as B300 and RTX PRO 6000 and distributed across the U.S., Europe, and India. When a particular card is difficult to source, partitioned capacity can provide the same memory class through slices of newer cards.

Storage is part of the same infrastructure problem; with VAST Data architecture backing Runpod's High Performance Storage, it rarely is. As models grow and datasets reach tens of terabytes, multi-node clusters can become constrained by training-data ingestion and checkpoint writes can limit GPU utilization.

VRAM planning tip

Start with the required memory class and workload shape. Fine-tuning, long-context inference, video upscaling, batch processing, and image generation place different demands on GPU memory.



Looking back at the last supply calls

The previous State of AI Compute report projected that H100 SXM and H200 supply would roughly double by mid-2026, while B200 supply would nearly quadruple.

At a glance

H100 SXM

Supply roughly doubled, closely matching the forecast

H200

Supply grew approximately 1.6x

B200

Supply tripled rather than nearly quadrupling

B200 growth fell short of the original projection, but B300 entered the fleet during the forecast period. Together, Runpod's B200 and B300 datacenter fleet is now roughly six times its size at the beginning of the year. Because the original forecast covered B200 alone, that combined figure provides context rather than a direct comparison with the forecast.

One expectation reversed. GPU price declines in 2025 gave way to published price increases across several providers in 2026. The change shows how quickly capacity economics can shift as demand, component availability, power, and infrastructure costs change.

Reviewing previous forecasts makes the uncertainty visible. The H100 SXM projection was close, while H200 and B200 grew less than projected. The difference between the B200 forecast and the eventual fleet mix also shows how quickly a newer GPU generation can change capacity planning.



Conclusion

Hardware supply tightened in 2026, but Runpod developers did not converge on one model or GPU. They quantized larger models, trained smaller specialists, combined open weights with frontier APIs, and chose memory classes around the workload.

These choices can reduce memory, latency, or idle-capacity requirements, although their effect on quality and total cost varies by deployment. Developers should define the workload first, then choose the model, precision, serving pattern, and hardware. As supply and prices change, those choices need to be tested again.