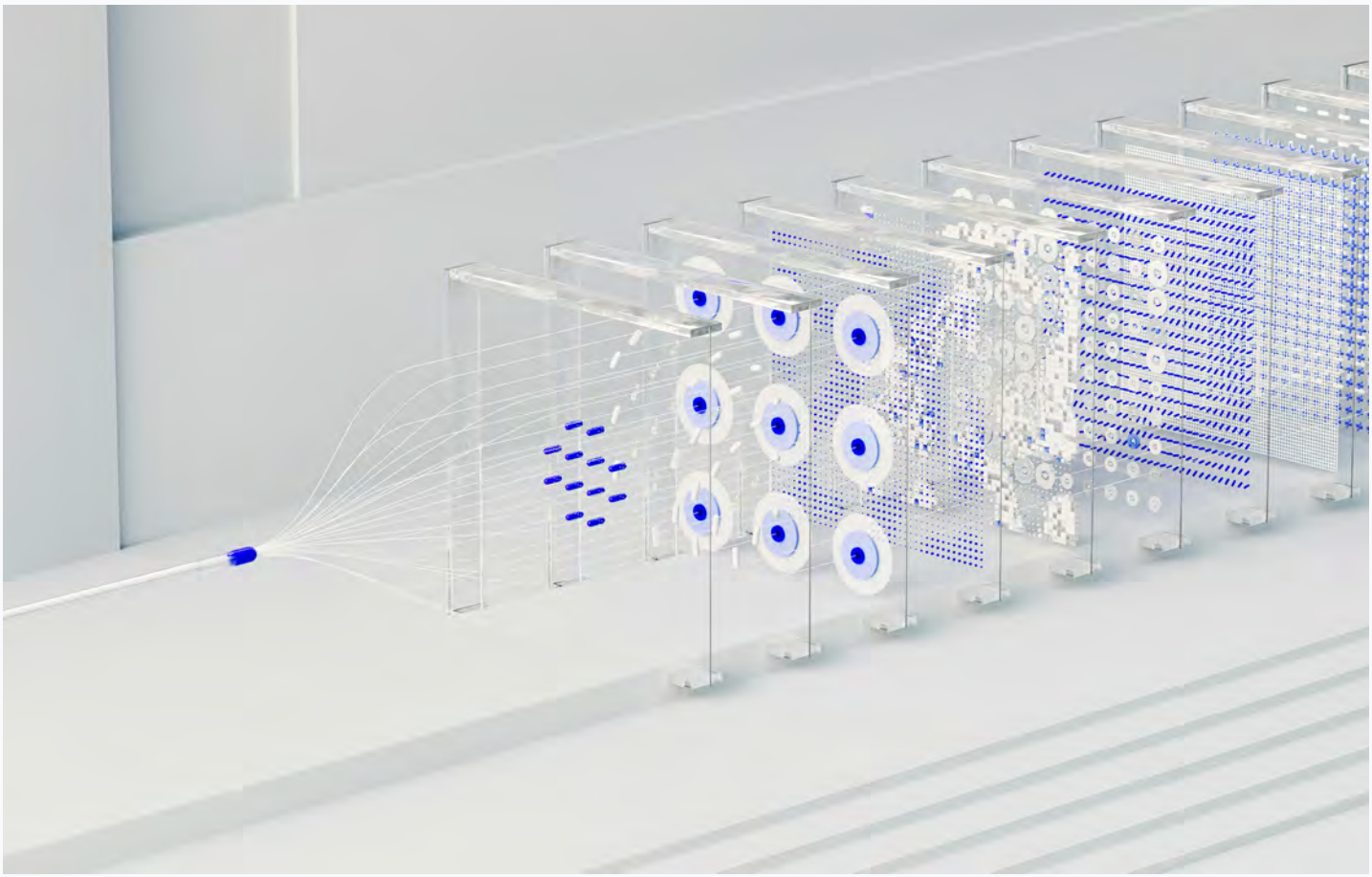
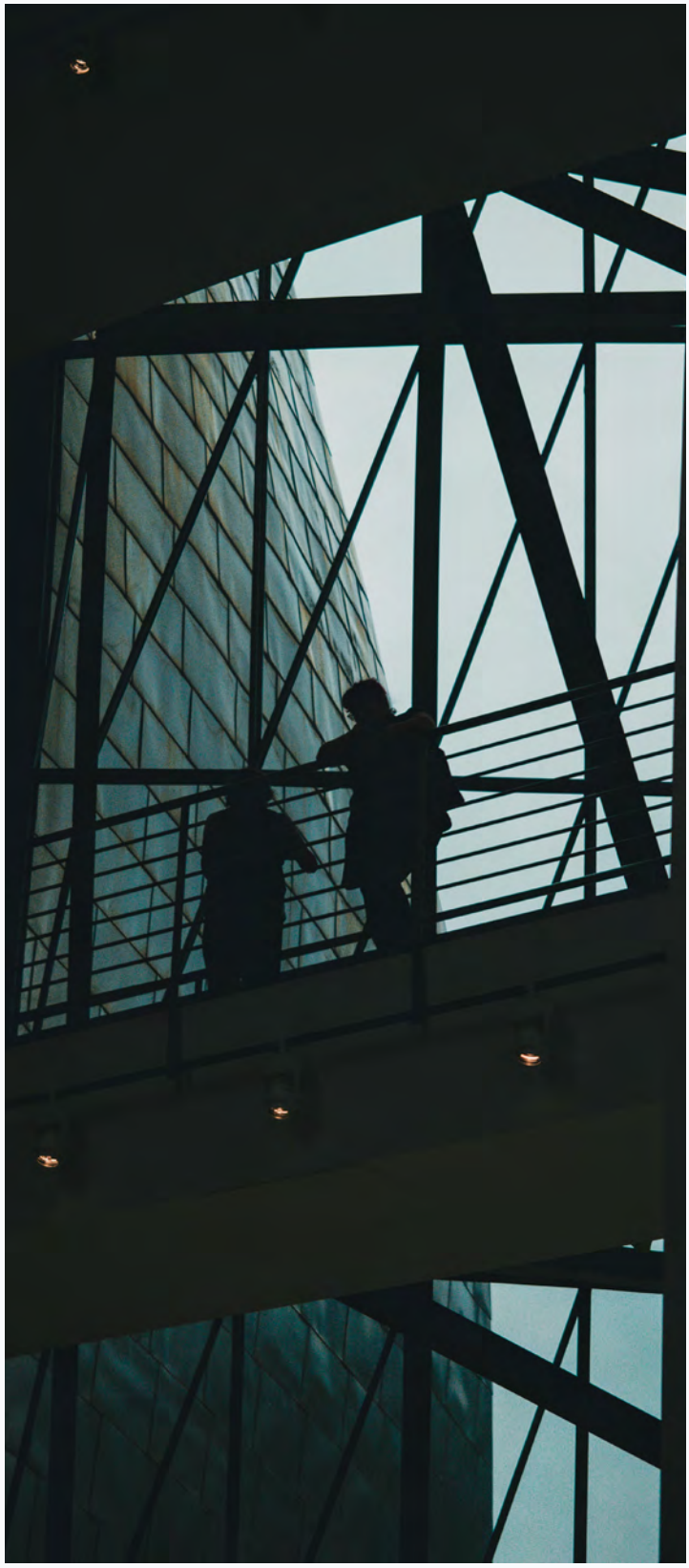
WHITEPAPER

The infrastructure moment in AI

Defining the Essential Cloud for AI



WWW.COREWEAVE.COM



01 A new foundation

Introduction

The flywheel of AI innovation is one of the most powerful creative forces of our time, and it's spinning ever faster. After decades of incremental AI progress, the first transformer models in 2017 signaled a paradigm shift. Suddenly, machines could grasp context, nuance, and meaning in ways that hinted at the next era of intelligence even if the full impact had yet to be felt.

The debut of GPT-3.5 in 2022 was the tipping point that catapulted AI into mainstream use and marked the start of frontier model competition. Today, the most advanced frontier models now double their effective time horizon every seven months (METR, 2025), but their effects cascade everywhere, including the products your teams create or adopt, the costs you manage, and the competitive pressures you face.

The reality is relentless. You're either accelerating innovation or falling behind. And the underlying foundation isn't a backdrop for this landscape, it represents your best opportunity for competitive advantage. Leaders face a critical fork in the road: stick with general-purpose clouds tuned for web apps and IT, or move to what the AI era demands—a purpose-built AI cloud engineered for massive training runs, low-latency inference, and rapid iteration.

PROOF IN PRODUCTION

OpenAI frontier model acceleration

CoreWeave AI Cloud is the networking backbone for GPT-5 model routing and orchestration across multiple clouds.

At AI scale, every layer of the stack sets the pace. Consider biology benchmarks where LLMs already match or surpass human experts, but only when training, inference, and tooling are optimized end to end (Justen et al., 2025). Hardware-software integration delivers parallel benefits, driving major gains in throughput and efficiency.

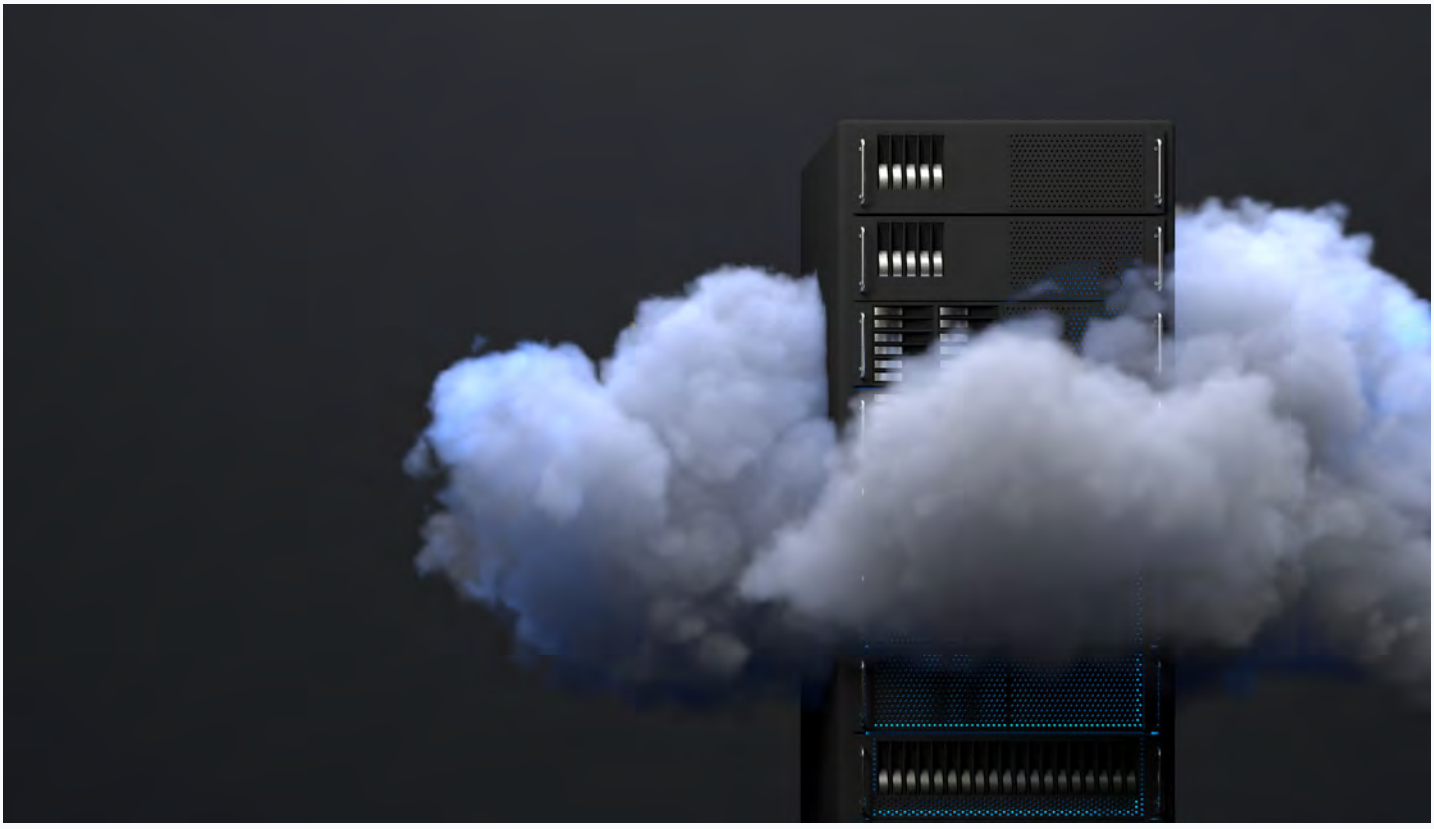
For pioneers, the takeaway is unmistakable: Breakthroughs demand a cloud built to deliver, scale, and sustain them. This paper introduces the purpose-built AI Cloud, a new foundation created not just to compete with hyperscale, but designed to replace it as the force multiplier for AI innovation.

02 Accelerating demand

Incredible opportunities and pressures

AI breakthroughs are everywhere, and they've unleashed a surge in demand as industries race to operationalize AI, straining the economics and availability of compute worldwide.

While many business leaders are debating whether AI is overhyped and worth the investment, the real pioneers don't have that luxury. They know the question isn't if AI will reshape their industry, but how fast—and they also know that every misstep risks lost opportunities and falling behind. AI is set to strain every layer of the IT stack, from hardware to data pipelines to governance, and the signals couldn't be clearer.



02 Accelerating demand

Surging demand

From chatbots to assistants, LLMs are now in the hands of millions of consumers while enterprises rush to embed them into core systems. That one-two punch is driving unprecedented strain on the underlying hardware, with global data center power demand expected to soar **165% by 2030** (Goldman Sachs, 2025).

Intensifying enterprise adoption

88% of enterprises have enhanced their ability to deploy AI at scale, strongly signaling that adoption is moving from pilots toward core infrastructure (Domino 2025 REVelate report).

Market paradox

The numbers are staggering and often seem unreal. AI is estimated to underpin $16 trillion of global GDP, with direct revenues projected to exceed $1 trillion by 2030 (IDCA, 2025). Yet the path to ROI is often murky, with both strategy and technology integral to success.

Together, these realities point to a simple truth: The foundational decisions leaders make now will define tomorrow's winners. As AI workloads multiply, the gaps in power, hardware supply, and scaling efficiency will only widen. Quick fixes and temporary workarounds may ease immediate pressures, but they can't deliver the sustained velocity, scalability, and cost efficiency that AI demands. Whether they're pioneering new AI products or modernizing core operations, leaders need a cloud that's engineered for AI from the ground up.

PROOF IN PRODUCTION

Toyota advances autonomous systems safety

Toyota worked with Weights & Biases by CoreWeave to automate driving-log bug classification, dramatically accelerating error detection and improving classification accuracy.

02 Accelerating demand

Market forces shaping the AI Cloud

Lack of talent:
New skill sets are in high demand and short supply; every company will need to reimagine roles and build new pipelines.

Sustainability concerns:
The compute boom's climate liability is growing, with pressure mounting on power, rare earths, and e-waste.

Challenging ethics:
Governance lags behind innovation, raising urgent questions about fairness, bias, and responsibility.

Geopolitical focus:
AI has become a national priority as sovereign governments pour billions into secure, regional ecosystems.

03 Pioneer principles

Principles for pioneers of the AI era

Hyperscalers enabled a generation of cloud innovation, but they were built in different times for different needs. At AI's magnitude, those environments create friction. AI pioneers need maximum velocity, and that velocity is defined by a new set of principles.

03 Pioneer principles

PRINCIPLE 1

Harnessing AI's power is the only path to leadership

This is not a wait-and-see moment. Those who innovate first will define the trajectory of the marketplace for the next decade and beyond.

PRINCIPLE 2

AI isn't just an accelerator, it's the chance to reinvent

AI will transform established markets and birth new ones. Enterprises that see AI only as a faster engine will miss the point; those who reimagine what's possible hold the power to redefine entire industries.

PRINCIPLE 3

Modern breakthroughs won't come from legacy mindsets or approaches

The scale and speed of AI demand new ways of thinking and organizing. Hierarchies built for incremental progress stall in the face of trillion-parameter models and dynamic ecosystems of models, agents, and humans. Pioneers need fresh foundations, both cultural and technical, to keep moving forward at speed.

PRINCIPLE 4

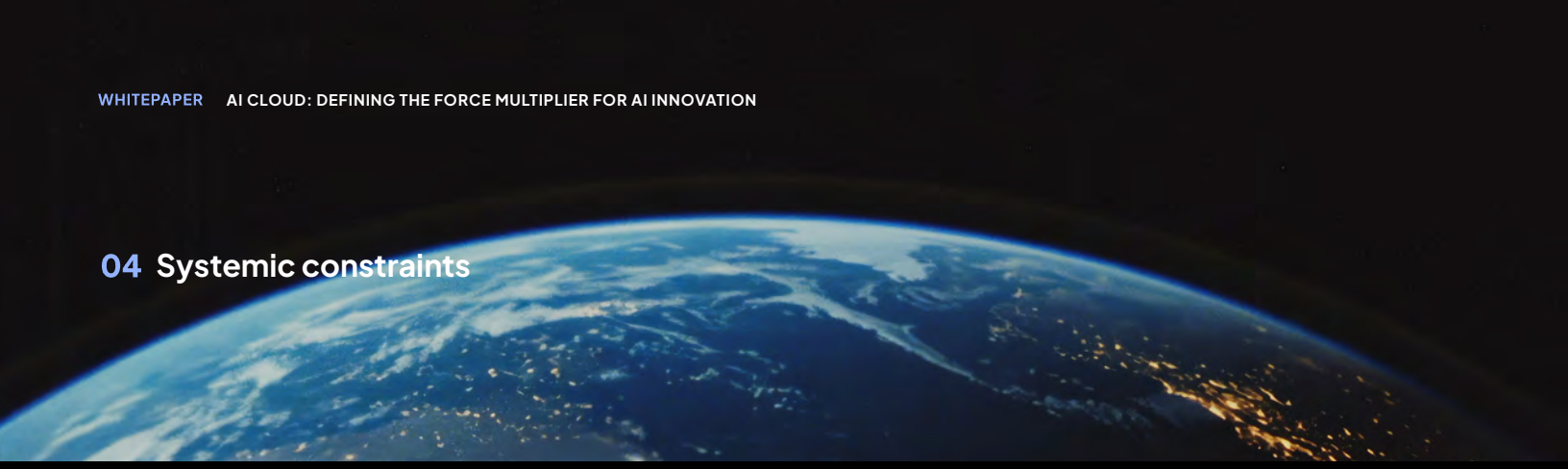
Learning and iterating at light speed are table stakes

In AI, extreme velocity means survival. Pioneers need platforms that let them experiment, fail, and scale without friction. First-to-market means relevance, recognition, and better ROI.

PRINCIPLE 5

The right IT architecture is the ultimate governor of innovation

Even the best models and teams hit a ceiling if the foundation can't support their requirements. With AI at scale, the cloud determines whether breakthroughs compound into lasting advantage or collapse under the weight of limiting bottlenecks.



04 Systemic constraints

Extreme scale and speed create massive challenges

Every era-defining shift comes with potential friction. When it comes to AI, those obstacles arrive fast and hit hard, turning technical challenges into market-shaping, innovation-defining constraints. They're not edge cases, they're the actual fault lines that determine who breaks through and who stalls out.

Just consider how a few of the most pressing challenges reveal the demand for a new model for building, deploying, and scaling AI.

> Vision without action is a daydream. Action without vision is a nightmare.
>
> — Unknown

04 Systemic constraints

Conventional playbooks collapse when scale and complexity collide

AI changes how code is built, deployed, and run, forcing teams to reengineer DevOps, CI/CD, and infrastructure patterns built for traditional software. Building agents, applications, and models requires purpose-built, integrated tools that are still maturing. Talent is scarce, observability is fragmented without unified dashboards, and conventional hierarchies falter in the fluid world of non-deterministic systems.

Proof-of-concept purgatory is real

Process breakdowns show up in the bottom line. Projects stall moving from POC to revenue-grade production as real-world data adds friction and ecosystem shifts make models harder to stabilize. ROI pressure at every funding gate forces shortcuts, leaving too much potential business impact unrealized. In an ecosystem that shifts rapidly, stabilizing models is a constant challenge, and waiting for it to settle means falling behind.

Without innovative thinking, governance is a drag

Every model requires traceability and audit trails baked in. AI safety and governance for protecting against everything from bias and toxicity to hallucinations and prompt attacks are not edge issues, they're mission-critical challenges. Yet expertise in AI safety, security, and compliance is thin, leaving organizations underprepared for escalating regulatory and risk demands. Instead of instilling confidence, governance too often slows velocity at the very moment when innovation demands speed.

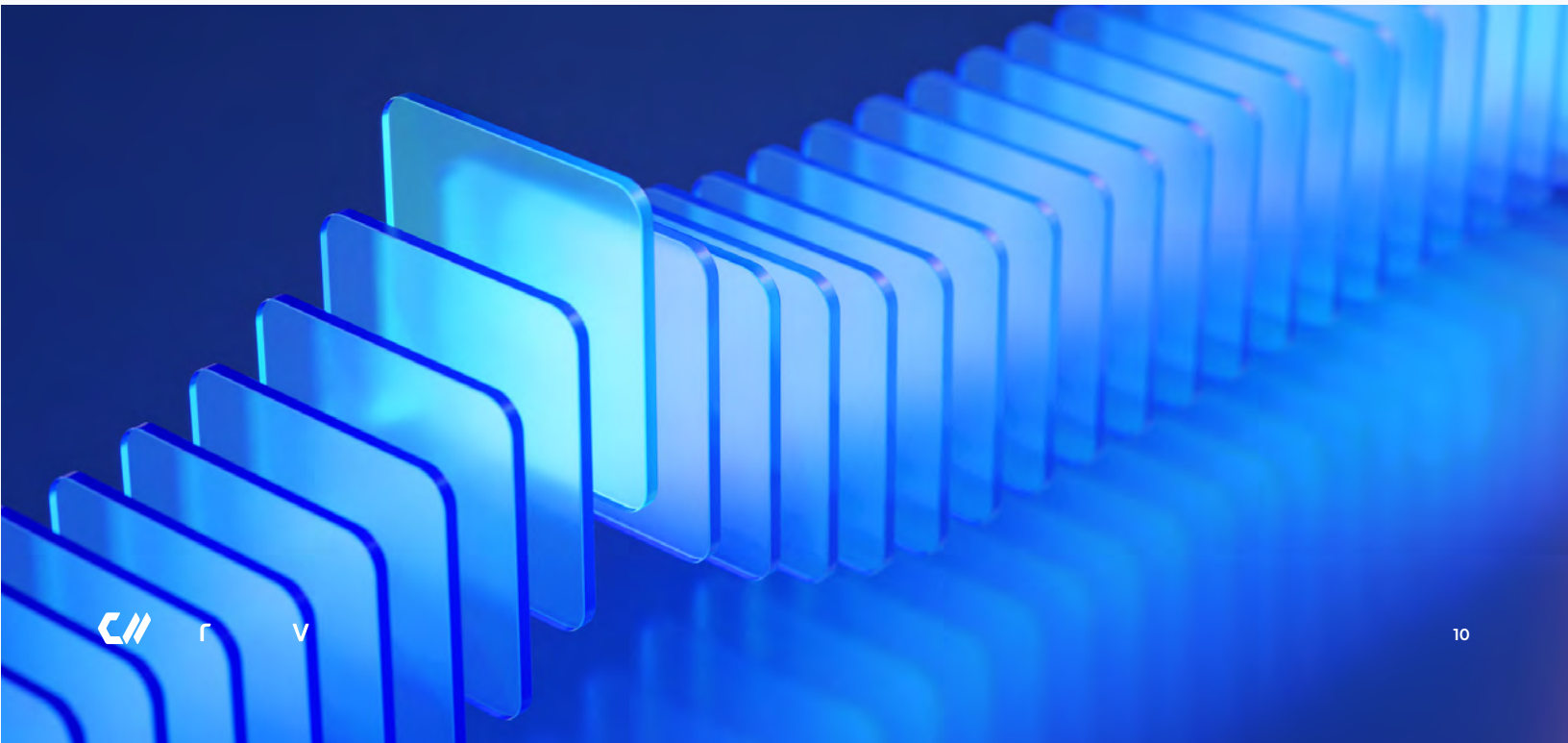
04 Systemic constraints

Costs throttle velocity

From limited GPU supply and long lead times that delay training schedules to price swings and traffic spikes that wreck cost forecasts, AI at scale is full of budgetary surprises. The rising complexity of models and agentic applications steadily drives costs higher, while balancing efficiency with tight SLAs demands constant tuning across hardware, models, and networks. Teams must justify TCO premiums with rapid, demonstrable performance gains, even as thousands of required iterations outpace the capabilities of most tools. Training-run crashes and inference failures only compound the pain, burning time, compute, and credibility. For pioneers, runaway cost isn't just a financial drag; in AI, velocity and survival go hand in hand.

Lock-in and data gravity undercut strategic flexibility

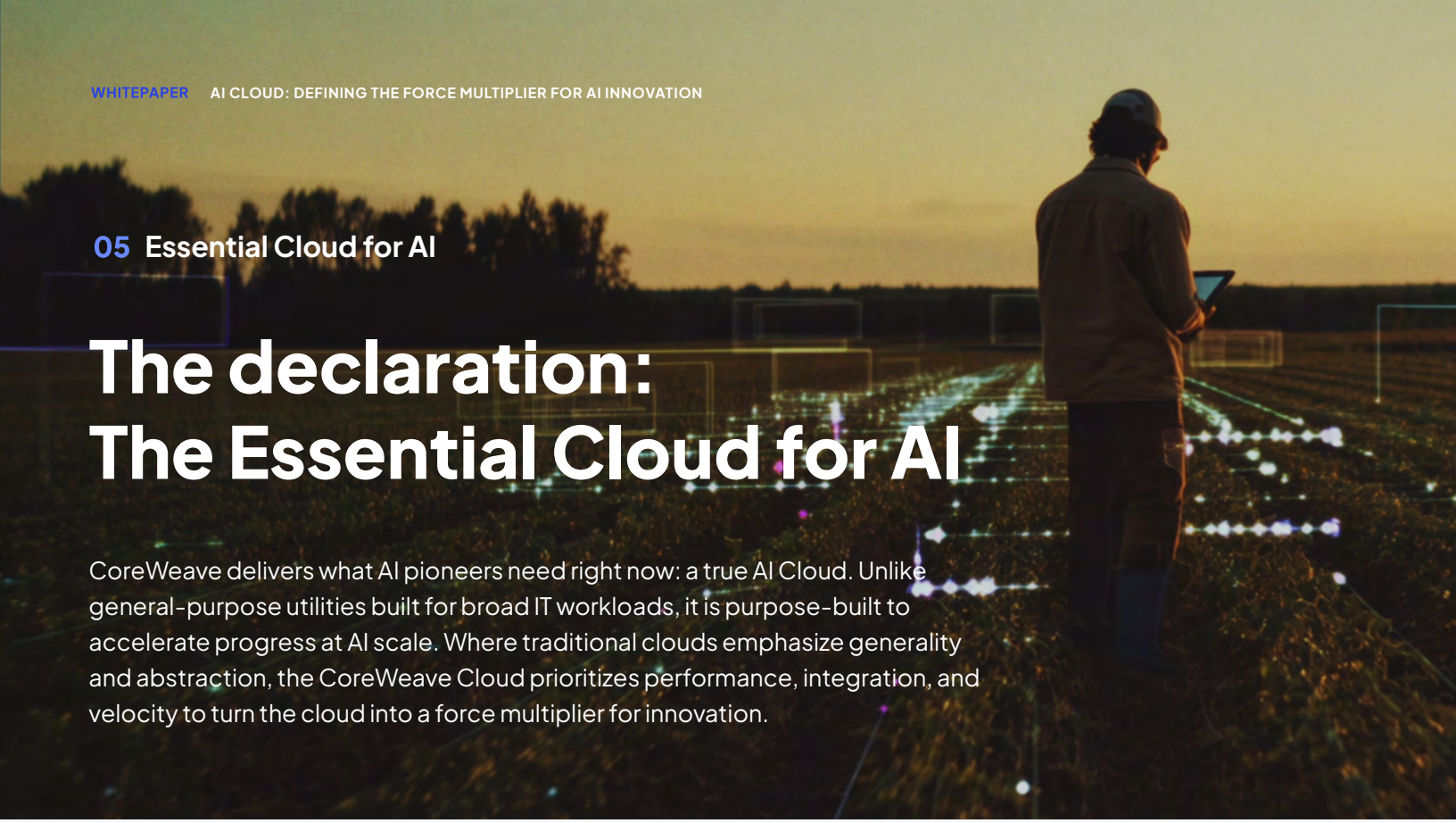
Hyperscaler and SaaS ecosystems lock margins and IP behind proprietary APIs, piling on high egress costs and long migrations. That lock-in stalls experiments and burns cycles at the very moment when flexibility to test, pivot, and negotiate pricing is paramount. And when proprietary ecosystems and data gravity dictate your roadmap, innovation slows to a crawl.



04 Systemic constraints

Why AI demands a different cloud approach

From the hyperscale model	To AI Cloud
Built for broad IT workloads	Built for high-performance control, observability, and optimization at every layer
Treats compute as generic, virtualized units	Exposes specialized hardware, including GPUs, liquid cooling, and high-speed interconnects directly for AI workloads
Separates stack layers to minimize risk	Integrates stack layers for efficiency, cost optimization, and performance
Incremental compute and architectural upgrades	Moves at the pace of AI innovation, rapidly adopting next-gen architectures
Built on multi-tenant, latency-tolerant networks	Low-latency, high-throughput pipelines where microseconds matter
Static input/output steps for data	Data flows dynamically and securely to where it's needed, fueling faster, more accurate outcomes



05 Essential Cloud for AI

The declaration: The Essential Cloud for AI

CoreWeave delivers what AI pioneers need right now: a true AI Cloud. Unlike general-purpose utilities built for broad IT workloads, it is purpose-built to accelerate progress at AI scale. Where traditional clouds emphasize generality and abstraction, the CoreWeave Cloud prioritizes performance, integration, and velocity to turn the cloud into a force multiplier for innovation.

Defining the next-gen approach

Conventional clouds treat compute as interchangeable and separate layers of the stack to minimize risk. That model works for web apps and transactional systems, but it falters with AI. Training, fine-tuning, inference, and agentic applications demand more direct access, control, and efficiency than generic infrastructure can deliver.

CoreWeave designed the CoreWeave Cloud to maximize performance and control across the entire lifecycle. Specialized hardware, such as bare-metal GPUs, liquid cooling, and high-speed interconnects, is made directly available to accelerate AI workloads.

Orchestration, scheduling, and monitoring operate as a single system of record (Mission Control for AI), backed by deep observability that runs all the way down to the chip. This integrated approach helps teams anticipate issues, optimize cost and performance, and keep critical workloads on track.

The result is a cloud that is fully optimized for the pace of AI innovation. Instead of incremental upgrades and generic scaling, it delivers low-latency, high-throughput pipelines where microseconds can determine outcomes. It moves data safely and efficiently to where it's needed, and balances the competing demands of scientists, developers, and cloud managers. In short, it's the foundation that AI pioneers need to unlock unmatched pace, maximum performance, and transformative partnership.

05 Essential Cloud for AI

Unmatched pace

CoreWeave Cloud turns speed into impact. By eliminating queues and streamlining every stage of AI development, teams can start training and deploying immediately instead of waiting weeks. Faster iteration compounds into shorter time to market and significant cost savings, whether you're scaling across dozens of GPUs or thousands.

Up to 25%
more FLOPs per GPU per hour, lowering cost per experiment

1M+ data points
ingested per second for faster data-to-model cycles

First
to bring the latest GPUs to market, accelerating performance leaps

Maximum performance

Built on failure-tolerant hardware, CoreWeave Cloud delivers the efficiency and reliability needed to handle everything from large-scale fine-tuning to high-throughput inference. Performance is optimized at every level, from low-latency agent orchestration for massive parallelism to proactive tuning that helps teams maximize utilization.

Up to 96% goodput
ensuring workloads run at peak speed without wasted cycles

20% higher MFU
(model FLOPS utilization), unlocking more performance and faster throughput from every GPU hour

100,000+ metrics
tracked per run for proactive optimization

250,000+ GPUs
proven in production, demonstrating performance at enterprise scale

05 Essential Cloud for AI

Transformative partnership

CoreWeave AI specialists amplify what AI teams and models can achieve by bringing deep expertise and direct-to-expert support across the AI lifecycle. With 24/7 direct-to-expert support included at no additional cost, CoreWeave resolves issues in hours, not days. And with full interoperability, CoreWeave AI Cloud gives teams the freedom to build on their terms so they can stay focused on tackling their biggest challenges and pursuing breakthroughs.

Direct-to-expert support, 24/7

Hours, not days, to resolution

Trusted by 1M+ practitioners

PROOF IN PRODUCTION

Mistral speeds time to market

Dedicated CoreWeave hardware and around-the-clock engineering support delivered 2.5× faster training cycles and fewer interruptions.

06 Engineered for scale

Inside the AI Cloud stack

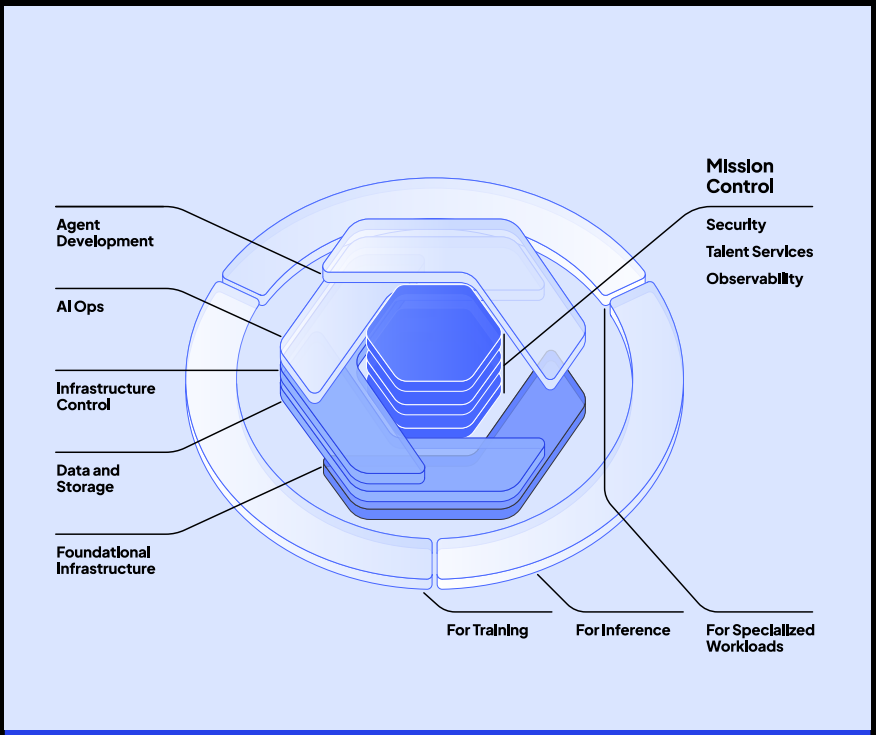
Efficiently delivering breakthroughs at AI scale takes far more than racks of GPUs. It takes a cloud designed from first principles, with AI-native optimization built into every layer. Where conventional clouds separate services behind layers of abstraction, CoreWeave integrates them into a system tuned end to end for training, fine-tuning, inference, and agentic applications.

Foundational technology

Force multiplication starts with a cloud purpose-built for AI: direct-to-GPU systems with bare metal performance, liquid cooling, high-speed interconnects, and failure-tolerant design. CoreWeave is consistently first-to-market with the latest GPUs, giving pioneers early access to performance leaps that otherwise take months to reach.

Data and storage

Breakthroughs rely on data that moves without friction. CoreWeave delivers distributed AI-native object storage, vector databases, and high-throughput pipelines designed for rapid data-to-model cycles, so training sets move quickly, inference stays low latency, and experiments scale seamlessly.



The CoreWeave stack includes everything AI pioneers need, and nothing they don't.

06 Engineered for scale

Cloud control

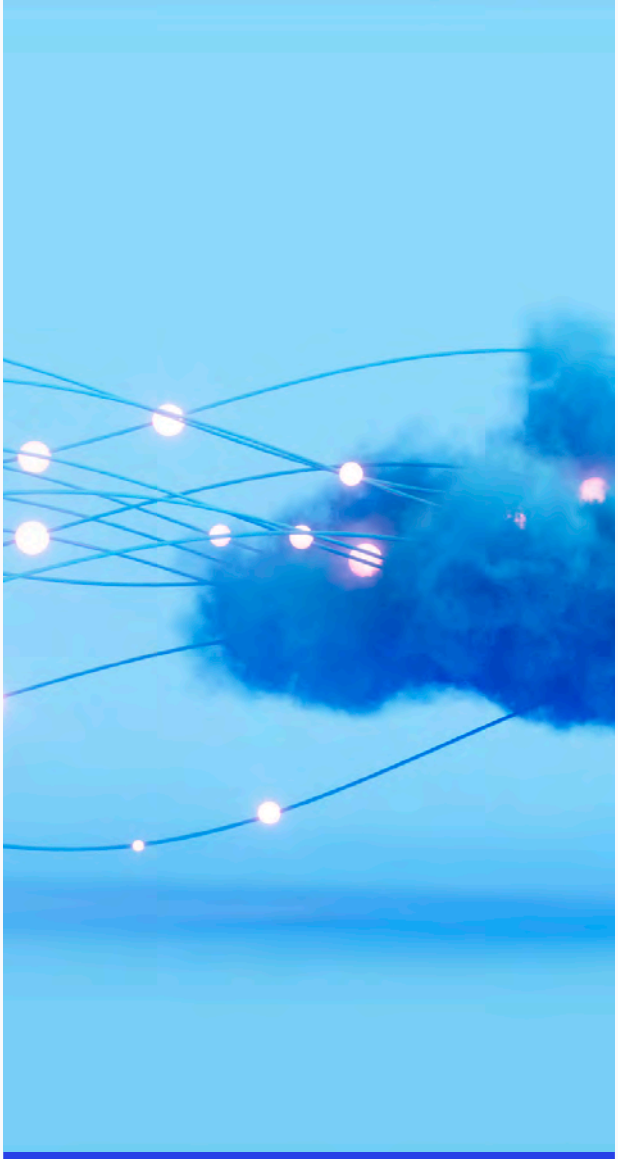
Velocity requires unified control and deep observability. Where conventional clouds force teams to juggle separate tools for orchestration, scheduling, and scaling, CoreWeave unifies them in Mission Control. This single system of record reconciles the competing demands of scientists, developers, and IT managers while surfacing 100,000+ metrics per run. With observability woven all the way down to the chip, teams can anticipate issues, optimize cost and performance, and keep workloads on track.

AIOps

Performance compounds when models are optimized in real time. CoreWeave embeds tools for model serving, scaling, and fine-tuning directly into the workflow. With observability baked in, teams can track model performance continuously and make adjustments on the fly.

Agent development

Innovation peaks in agentic workflows, where thousands of iterations unlock new frontiers. CoreWeave provides the foundation and tooling to develop, test, and deploy agentic workflows complete with built-in guardrails for safety and reliability. Teams can move from prototype to production quickly, backed by direct-to-expert support.



Cross-stack accelerators

Three elements strengthen every layer of the CoreWeave Cloud: **security** that wraps every interaction, **observability** that shows impact in real time, and **talent services** that help teams chart new solutions rather than falling back on old playbooks.

Collectively, these strategic choices and purpose-built capabilities place CoreWeave in a class of its own: a true AI Cloud for pioneers determined to outpace their competition.

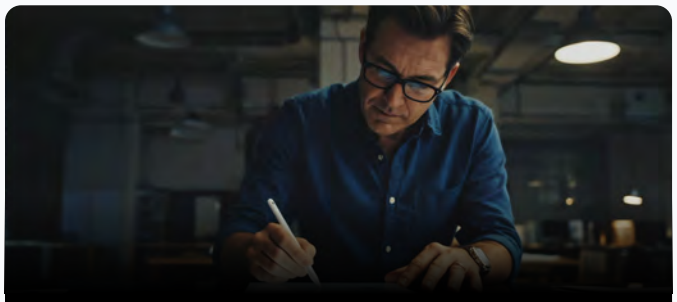
07 Performance without limits

Built to accelerate every AI workload

Large-scale AI training

Model training • fine-tuning • reinforcement learning

CoreWeave accelerates model development with bare-metal GPU performance, high-throughput AI object storage, and deeply integrated AIOps that simplify workflows, so teams train faster, more efficiently, and at lower cost.



CASE STUDY:

IBM accelerates research breakthroughs

Deploying CoreWeave with NVIDIA GB200 NVL72 GPUs, IBM accelerated MLPerf training speeds by 80%+, giving researchers one of the world's fastest AI environments.

Inference and agentic applications

RAG • agentic orchestration • inference • AI application dev

For agentic and inference workloads, CoreWeave's developer environment, powered by Weights & Biases, gives teams industry-leading tools to build, test, and deploy with AI instead of wrestling with hardware. Fully integrated into CoreWeave's bare-metal cloud, it delivers the low-latency, high-throughput performance required for real-time inference and agentic applications at scale.

CASE STUDY:

Carbon Re lowers CO_2 emissions and costs

With Weights & Biases Sweeps and Models, Carbon Re's Delta Zero Cement optimizes kiln operations in real time, lowering fuel costs and cutting CO_2 emissions via a centralized model registry.

07 Performance without limits

Specialized workloads

Computer vision • VFX rendering • simulation, physics, digital twins • scientific computing and HPC

With bare-metal GPU and CPU performance, high-bandwidth networking, and scalable storage, teams accelerate research, design, and production pipelines, unlocking capabilities that traditional cloud environments struggle to support.

CoreWeave is the only AI Cloud provider to receive the SemiAnalysis Platinum ClusterMAX™ Rating among engineers, developers, and enterprises.

08 The CoreWeave Effect

Conclusion: the CoreWeave Effect

From frontier labs putting intelligent models in people's hands to changemakers reinventing mission-critical functions, there's a clear need for a foundation that propels progress. CoreWeave is that force multiplier, turning vision and ambition into advantage.

Momentum

Move faster, learn faster, ship faster. With up to 20% higher MFU than industry peers, CoreWeave turns wasted cycles into acceleration.

Jobs finish sooner, models ship faster, and teams get to market ahead of the curve.

Magnitude

Scale without limits to ensure ideas reach their full potential.

With up to 25% more FLOPs per GPU per hour and early access to the latest GPUs, every dollar of GPU spend goes further. That efficiency compounds at scale, delivering more progress per dollar.

Mastery

CoreWeave earned the only platinum rating from SemiAnalysis, and today serves the most advanced AI labs in the world, including OpenAI, Cohere, and Mistral.

Operate with confidence. End-to-end observability, built-in governance, and direct-to-expert support turn risk into reliability.

When pioneers meet the first true AI Cloud, progress compounds. That's what we call the CoreWeave Effect. Changemakers across industries are already building faster, scaling further, and unlocking possibilities that traditional models could never support.

09 Next steps

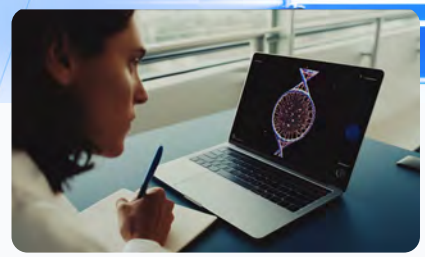
Build your AI leadership momentum

CASE STUDY:

Mistral AI unlocks 2.5× faster training speeds

Learn how Mistral AI boosted model training speed by leveraging the NVIDIA GB200 NVL72 platform on CoreWeave Cloud.

READ THE CASE STUDY

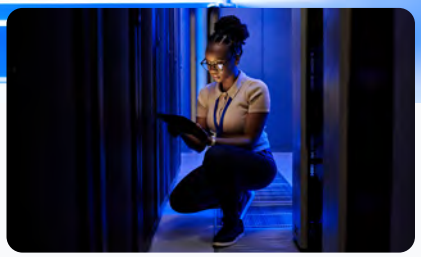


BENCHMARK REPORT:

MLPerf V5.0 results on training and inference

See why the CoreWeave Cloud leads MLPerf benchmarks across both training and inference performance.

READ THE REPORT



DIVE DEEPER:

Why CoreWeave Cloud?

CoreWeave Cloud is the force multiplier for AI development, offering the pace, performance, and partnership you need to accelerate AI innovation.

TAKE THE NEXT STEP