CASE STUDY

COHERE x COREWEAVE

Cohere Case Study

INDUSTRY	HEADQUARTERS	USE CASES
Artificial intelligence	Toronto, Canada	Agentic AI

RESULTS: SPEED AT SCALE

3x
Faster training[1]

Up to 7 GB/s
Throughput/GPU[2]

Challenge

Cohere is building the next generation of AI for the enterprise, delivering production-ready LLMs and agentic AI systems for real-world applications.

For its recent pioneering project, Cohere needed a compute and storage foundation powerful enough to support North, their secure agentic AI platform for enterprises, while maintaining the agility and efficiency that define their approach to innovation.

To achieve this, Cohere required a cloud partner that could deliver cutting-edge compute performance, high-throughput AI storage, and the operational reliability needed to run next-generation agentic workloads at scale.

Core Needs List

— Massively scalable compute to support rapidly rising AI workloads and training cycles

— High-performance storage to overcome cloud object-storage limitations at scale

— A proactive partner capable of deploying NVIDIA GB200 NVL72 early and reliably

> We knew we needed a partner who could provide the compute and newest processing technologies, and also keep pace with our team.

Autumn Moulder
VP of Engineering, Cohere

Solution

Cohere partnered with CoreWeave to deploy one of the industry's first NVIDIA GB200 NVL72 clusters in production. Having collaborated across multiple GPU generations—including A100, H100, and GH200—they trusted CoreWeave's ability to operationalize new hardware early and at scale.

This deployment combined high-performance compute with CoreWeave AI Object Storage to accelerate data movement, simplify replication across clouds, and reduce costs.

01 **Early access to NVIDIA GB200 NVL72**
CoreWeave provided Cohere with early GB200 capacity for testing, enabling them to validate workflows and identify issues before production deployment. This accelerated Cohere's development process for North and future large-scale models.

02 **Proactive hardware health management**
CoreWeave's automated rack-health monitoring (e.g., validating cluster node health) surfaced issues earlier than other cloud providers, reducing operational risk for Cohere during early adoption.

03 **Operational excellence with CoreWeave Kubernetes Service (CKS)**
CKS delivered reliable operations with built-in observability tooling and declarative nodepool management via Infrastructure As Code, requiring minimal intervention from Cohere's team.

04 **ARM64 validation through GH200 nodes**
GH200 availability allowed Cohere to test ARM64 compatibility with their software stack, uncovering surprises and resolving them before full GB200 rollout.

05 **High-performance AI Object Storage**
CoreWeave AI Object Storage provides sustained throughput of up to 7 GB/s per GPU, to enable multi-region and multicloud unified datasets, reduced replication friction, and introduced cost-efficient tiering.

> "Because we had access to compute early, we were able to optimize for speed and efficiency. When it came time to train models for North, our team was able to focus on training iterations to bring our enterprise customers efficient and secure agentic AI."
>
> **Autumn Moulder**
> *VP of Engineering, Cohere*

UNLEASH AI'S POTENTIAL

Outcomes

The deployment of NVIDIA GB200 NVL72 on CoreWeave enabled Cohere to accelerate experimentation cycles, iterate more quickly on large-scale models, and bring North to its customers faster. With higher performance and more consistent throughput, Cohere can sustain rapid innovation while strengthening its position as a leader in enterprise-ready, agentic AI.

3x Faster training performance at scale

Cohere achieved up to 3× higher training performance for 100B-parameter models compared to previous-generation Hopper GPUs—even before Blackwell-specific optimizations. This uplift allows their team to train more frequently, shorten iteration loops, and accelerate innovation velocity.

High-throughput multicloud data access

Cohere realized consistently high storage throughput up to 7 GB/s/GPU across clouds and regions, powered by CoreWeave AI Object Storage. This enabled a unified dataset architecture without the performance penalties or replication friction common in traditional cloud object storage.

Faster time to market on next-gen AI

With early access to GH200 and GB200 hardware, proactive cluster health tools, and seamless CKS operations, Cohere moved from prototype to production faster. Shorter training cycles and reduced overhead drive meaningful ROI and enable earlier delivery of new AI capabilities.

1 "Thousands of NVIDIA Grace Blackwell GPUs Now Live at CoreWeave, Propelling Development for AI Pioneers," NVIDIA, April 15, 2025.
2 "CAIOS Achieves 7+ GB/s per GPU on NVIDIA Blackwell Ultra," CoreWeave, September 22, 2025.

Scale confidently on CoreWeave

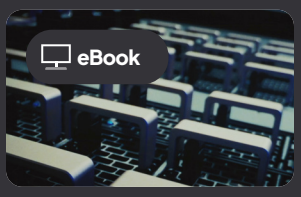
Learn how CoreWeave helps enterprises run, scale, and optimize next-generation AI workloads.

Contact Us

Rise of the AI Cloud

Discover what drives global AI innovation.

Watch now →



Smarter Compute Spend

5 things to consider before you buy infra.

Download →

Storage Without Limits

Explore the latest in AI Object Storage.

Read more →