

IBM boosts performance by 80%



INDUSTRY
Artificial intelligence

HEADQUARTERS
New York, NY

USE CASES
Model training

Explore AI Cloud

Learn how CoreWeave helps organizations scale their most ambitious AI innovations.

[See it in action](#)

DELIVER RESULTS AT SCALE

70B LLM

Parameters

1.3B

Search queries per month

11M+

AI-generated answers daily

“

“We knew that we needed more capacity, and we knew that we needed faster infrastructure in order to meet the roadmap for the Granite™ models.”

Danny Barnett

VP of Emerging
Technology Engineering,
IBM Research

Challenge

IBM has long pushed the boundaries of technological innovation. As an enterprise-grade tech giant, their teams have consistently been on the forefront of progress—from mainframe computing to quantum computing, as well as groundbreaking AI technologies like watsonx™ and their latest [Granite™ family of models](#).

With IBM® Granite™, IBM wanted to empower enterprise-level businesses with smaller, open source AI models that would give their clients enterprise-ready multimodal models. The goal? Deliver unparalleled performance across a wide range of enterprise tasks—from cybersecurity to Retrieval-Augmented Generation (RAG)—at up to 90% lower costs compared to larger frontier models.

To accelerate their roadmap for Granite™ models, IBM needed access to **greater capacity, scalability, and faster infrastructure**. By leveraging a cloud platform, purpose-built for AI, they sought to achieve a clearly defined set of benefits:

- **Offload infrastructure management tasks** that would allow their team to focus on faster, safer innovation, helping their clients get modern AI models to market faster.
- **Access a suite of solutions** that maintain GPU cluster resiliency, keeping jobs fully operational with greater performance and lowering total cost of ownership.
- **Maintain greater flexibility and agility** in building open, multi-solution environments that would allow IBM to migrate the critical tools they needed from previous AI infrastructure builds.

Solution

With IBM's extensive experience building solutions on-prem, they knew the type of challenges they would face in constructing and managing purpose-built AI infrastructure—and they understood the advantage of partnering with an experienced AI cloud provider to mitigate that complexity. When the time came to choose a partner, [IBM decided to work with CoreWeave](#) due to our platform's unique hyperspecialization in supporting Gen AI use cases.

IBM's teams were also confident that CoreWeave would help them tackle their biggest AI infrastructure challenges, thanks to CoreWeave's strong track record of **working with leading AI labs and enterprise giants**. They believed in CoreWeave's ability to provide the compute capacity at scale they required, based on our reputation for being consistently **first to market with supercomputing scale**.

CoreWeave worked closely with IBM to engineer a robust environment uniquely designed to meet their AI infrastructure requirements, including:

01 **NVIDIA GB200 NVL72:**

IBM gained access to several thousand GB200s, allowing them to build a supercomputer twice the size of their previous supercomputer, Blue Vela, which was already ranked among the top 10 largest in the world.

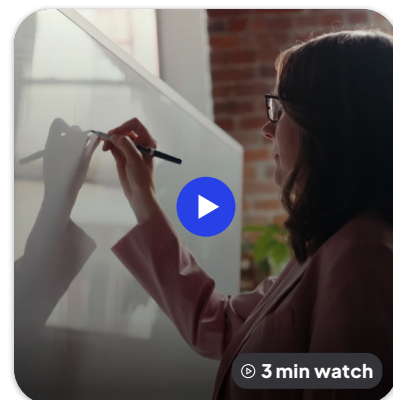
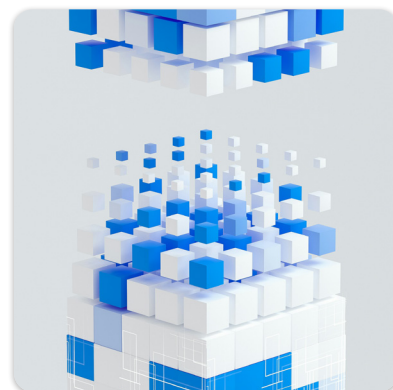
02 **CoreWeave Mission Control:**

CoreWeave's cluster management solution Mission Control automates the remediation of infrastructure challenges and enhances cluster resiliency, freeing IBM researchers from constant firefighting so they can stay focused on what they do best: pushing the boundaries of AI innovation.

03 **24/7 Support with Co-Creation:**

Consistent and open communication allows IBM and CoreWeave's teams to collaborate together and actively customize and co-create new environments. This level of trust and partnership enabled IBM and CoreWeave to be one of the first instances in the world to stand up NVIDIA GB200 NVL72 clusters.

A NEW FOUNDATION



AI Innovators in Action: IBM

IBM shares how CoreWeave pushed boundaries to deliver cutting-edge AI powered by the latest NVIDIA GPU's

UNLEASH AI'S POTENTIAL

1.8x faster training
Leader in AI innovation

24/7 support
Keeping pace with the AI industry

Up to **90%** lower costs
For end users

Results

By unlocking unprecedented speed and scale in AI innovation, IBM empowered their researchers to train Granite™ models with impressive performance at unparalleled speeds. Hillery Hunter explains, “With our partnership with CoreWeave, we are adding one of the world’s best-performing AI environments with these NVIDIA GB200 GPUs. This gives us yet another significant boost in productivity.”

IBM was able to leverage CoreWeave’s highly innovative AI cloud to achieve the following industry-changing results:



Groundbreaking MLPerf results

Greater than 80% faster speed and enhanced performance results on NVIDIA GB200 NVL72 impressed the industry and the world.



Marketplace impact

As an early adopter of the NVIDIA GB200 NVL72 cluster, IBM significantly accelerated the performance of its Granite™ models—enhancing client workloads and driving innovation across the AI marketplace.



A deep technical partnership with a hyperspecialized provider

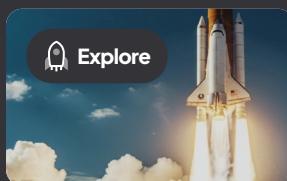
Leveraging shared knowledge and expertise, IBM and CoreWeave partnered together to create purpose-built environments featuring AI-optimized infrastructure that flexibly supports IBM’s needs for high-performance computing at scale.

This partnership not only accelerates Granite™ models to time-to-market, but also clearly demonstrates how organizations of any size, even established enterprises like IBM, can thrive by leaning into hyperspecialized infrastructure providers built for the AI era.

Take the next step with CoreWeave

Unleash the full potential of the most ambitious AI innovations with the CoreWeave Cloud.

[Contact Us](#)



Powering innovation

See how CoreWeave’s AI cloud delivers more.

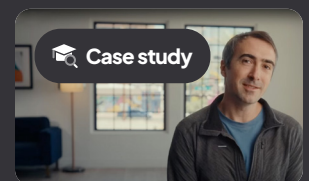
[Read more →](#)



Why CoreWeave Cloud

Learn why CoreWeave Cloud is the force multiplier.

[Learn more →](#)



Mistral AI

See how Mistral AI trained models 2.5x faster.

[Learn more →](#)