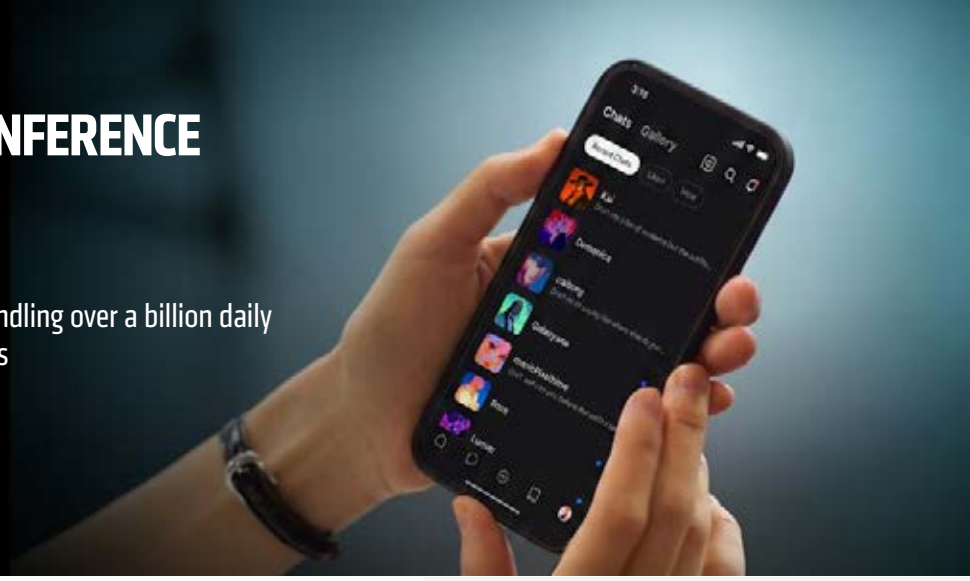
# CHARACTER.AI DOUBLES AI INFERENCE THROUGHPUT WITH AMD

CASE STUDY

Character.AI cuts its serving costs by 50% while handling over a billion daily queries using DigitalOcean with AMD Instinct™ GPUs

AMD × (character.ai)



Character.AI is home to tens of millions of AI-generated characters that users adore to the tune of more than a billion queries every day as they spin up stories, role-play scenarios, and have conversations with characters. These types of engagements create high-stakes technical challenges. Every millisecond of lag can cause characters to take too long to respond. Inference performance is the difference between an experience that is immersive and one that feels like waiting for a loading screen.

> **"AMD Instinct™ GPUs enabled a 2X improvement in production inference throughput and lowered our cost per token by 50 percent."**
> David Brinker, Senior Vice President, Character.AI

As the platform's global audience continues to grow, GPU infrastructure emerged as the company's single biggest operating cost. Character.AI needed a sustainable economic model for large-scale production inference. As technical limits were also becoming a barrier to its infrastructure roadmap, the company required a modern architecture that could handle technologies such as prefill decoding, disaggregated serving, and distributed KV cache. These depend on high-bandwidth connectivity, which its existing infrastructure lacked. To address the situation, Character.AI worked with DigitalOcean to deploy a new inference environment built on AMD Instinct™ MI300X and AMD Instinct™ MI325X GPUs.

## FINDING THE RIGHT GPU FOR MASSIVE MODELS

Evaluating new hardware required the Character.AI team to closely examine memory capacity and network fabrics. "We quickly realized that AMD Instinct GPUs would give us the fat, fast pipes between GPUs that our workload needs," says David Brinker, Senior Vice President, Character.AI. "Our assumptions about how its high-bandwidth memory would handle larger Mixture-of-Experts models and aggressive KV caching played out exactly as we'd hoped. Moving to DigitalOcean with AMD Instinct GPUs gave us the connectivity and hardware performance to deploy modern serving architectures at scale."

> **"AMD Instinct GPUs would give us the fat, fast pipes between GPUs that our workload needs."**
> David Brinker, Senior Vice President, Character.AI

**INDUSTRY**
Entertainment

**CHALLENGES**
Character.AI needed infrastructure that could support newer serving architectures, handle latency-sensitive inference at scale, and hold up under constant production demand

**SOLUTION**
Character.AI turned to AMD Instinct™ MI300X and MI325X GPUs in a new DigitalOcean cloud environment shaped by high-bandwidth connectivity and brought to life through close engineering collaboration

**RESULTS**
Character.AI doubled production inference throughput, cut cost per token by 50%, and gained room to serve more users and offer more advanced models

**AMD TECHNOLOGY AT A GLANCE**
AMD Instinct™ MI300 Series GPUs

*To keep AI conversations fast and immersive at massive scale, Character.AI built a new inference foundation with AMD Instinct™ GPUs.*

## MOVING TO AMD WHILE KEEPING FAMILIAR TOOLS

Because this was the first time Character.AI deployed AMD Instinct GPUs, the team had to port their existing workloads using the AMD ROCm™ software stack. Character.AI relies on the open-source vLLM framework for serving, so the move had to work across the serving stack, not just at the hardware layer.

In one production deployment built around the Qwen3-235B Instruct FP8 model, developers kept the workloads running smoothly by combining the AMD ROCm software stack, vLLM, and AMD AITER, a high-performance library of kernels to accelerate LLM training and inference on AMD Instinct GPUs. Native hardware support for FP8 execution on AMD Instinct MI325X GPUs kept the system on the fast path and eliminated the need to cast data types between memory and compute.

> **"DigitalOcean and AMD…allowed us to transition our serving stack without having to reinvent our approach."**
>
> David Brinker, Senior Vice President, Character.AI

"Bringing a new GPU architecture into production requires strong partnerships," says Brinker. "DigitalOcean and AMD helped optimize our model kernels and manage the infrastructure. This allowed us to transition our serving stack without having to reinvent our approach."

## PARTNERS ROLL UP THEIR SLEEVES TO DRIVE SUCCESS

Character.AI prioritized reliability, solution maturity, and partners' willingness to co-develop technical solutions. The team did not want a vendor that would simply hand over hardware and walk away. "We need partners who are committed to working through the challenges that come with a massive deployment," says Brinker.

Together, the teams analyzed bottlenecks, evaluated the available headroom, and built the exact Kubernetes manifest Character.AI needed to make its configuration work. The result is a managed Kubernetes platform that handles networking, gateway, and inner-cluster routing so workloads scale when put into production. Moving to a new data center also meant different network fabrics and software version requirements. "AMD and DigitalOcean sat down with us, dug into the bottlenecks, and helped us figure out the exact configurations we needed to fully unlock data parallelism," says a Character.AI executive. "That level of technical collaboration is indispensable."

> **"AMD Instinct GPUs give us room to serve more users on the same compute footprint…and grow our business."**
>
> David Brinker, Senior Vice President, Character.AI

The teams looked beyond standard tensor parallelism to instead deploy highly efficient setups that marry data, tensor, and expert parallelism across GPUs. They optimized parallelization strategies for the Mixture-of-Experts models employed by Character.AI to maximize queries per second while simultaneously meeting Character.AI's latency targets. "We challenged DigitalOcean and AMD to meet our strict latency constraints at scale," says a Character.AI executive. "Using AMD Instinct GPUs enabled a 2X improvement in production inference throughput and lowered our cost per token by 50 percent."

A Character.AI executive continues, "AMD Instinct GPUs give us room to serve more users on the same compute footprint, make more advanced models available to our community, and grow our business. For an AI-native company where inference is the operating budget, the efficiency gains made possible by AMD Instinct GPUs can equate to millions of dollars."

*Working with AMD and DigitalOcean, Character.AI optimized its serving stack without reinventing its development workflow.*

## BUILDING THE FUTURE OF CONSUMER-LED AI ENTERTAINMENT

The new inference foundation gives Character.AI the capacity it needs to start building its next generation of consumer AI products. The throughput and cost gains Character.AI has achieved are helping to make that product vision more practical to pursue.

> **"The AMD hardware roadmap gives us the confidence to keep extending the frontiers of multimodal content creation."**
>
> David Brinker, Senior Vice President, Character.AI

Character.AI is starting to evaluate the AMD Instinct MI350 and MI355 GPUs to leverage native FP4 support and further improve model inference efficiency. The team also anticipates benefiting from future improvements to memory and network switching that will continue to reduce bandwidth bottlenecks. "We are in the early stages of realizing what AI entertainment can become," says Brinker.

"AMD Instinct GPUs solved our immediate scale challenges, and the AMD hardware roadmap gives us the confidence to keep extending the frontiers of multimodal content creation."

*Higher inference throughput and lower serving costs help Character.AI advance the next generation of AI-powered entertainment.*

**EXPLORE AMD INSTINCT GPUS**
Sign up to receive AMD data center communications.

### ABOUT CHARACTER.AI

Character.AI is a Menlo Park company founded in 2021. Millions of people visit the platform each month to build stories, role-play scenarios, and have conversations with user-generated AI characters. Its creative tools include no-code character creation, real-time voice and video features such as Character Calls and AvatarFX, and an AI-native feed for sharing, remixing, and exploring AI-generated content. For more information, visit Character.AI.

### ABOUT AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics, and visualization technologies. Billions of people, leading Fortune 500 businesses, and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible. For more information about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn, and X pages.

### DISCLAIMERS



### COPYRIGHT NOTICE