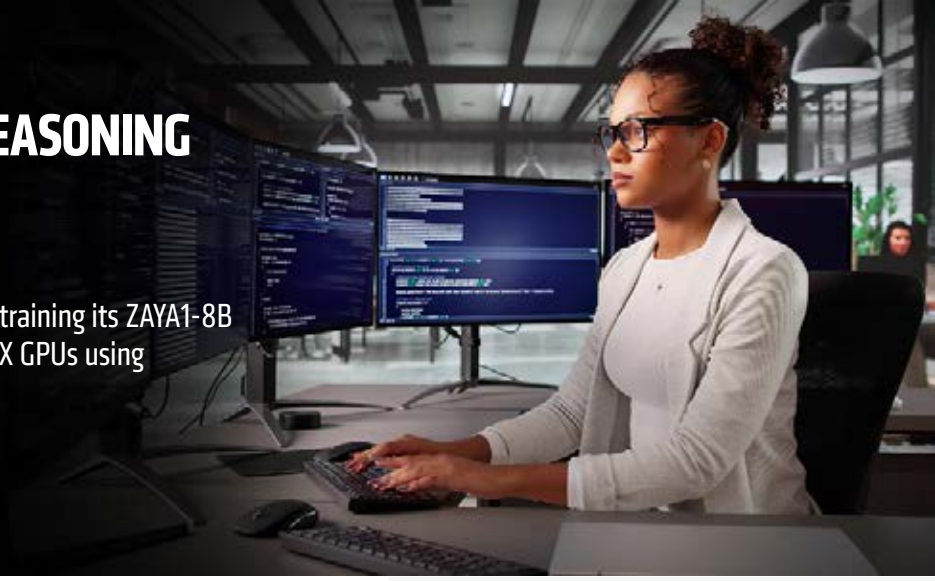
ZYPHRA TRAINS ADVANCED REASONING MODELS AT SCALE WITH AMD

CASE STUDY

Zyphra demonstrates full-stack AMD AI training by pretraining its ZAYA1-8B model from scratch across 1,536 AMD Instinct™ MI300X GPUs using AMD Pensando Pollara Networking

AMD × ZYPHRA



Zyphra is an open superintelligence research and product company based in San Francisco. Its mission is to ensure intelligence remains open, across models, infrastructure, and hardware, so organizations can develop and deploy AI on their own terms. Zyphra Research pushes the frontier of open intelligence, training multimodal models on heterogeneous compute. Zyphra Cloud delivers that work at scale with a full-stack AI platform built on AMD for long-horizon agents serving the advanced AI needs of developers, enterprises, and frontier hyperscalers. Achieving this mission requires models that can reason across long context windows, retain more of what users teach them, and keep learning over time.

> **"We trained about 20 trillion tokens on our AMD Instinct™ MI300X GPU cluster and achieved many uninterrupted 48-hour runs."**
>
> Beren Millidge, Co-founder and Chief Scientist, Zyphra

Zyphra's ambition for open superintelligence necessitated a scale-up of its training roadmap and platform. Their previous platform limited how much of a model could stay resident on a single accelerator or node, leading to more parallelism and communication overhead. Those constraints would only become harder to manage as training scaled. Zyphra chose a different direction, standardizing on an end-to-end AMD platform consisting of AMD Instinct™ MI300X GPUs, AMD ROCm™ software, and AMD Pensando™ Pollara 400 AI NICs.

TRAINING ADVANCED AI AT SCALE ON AMD

ZAYA1-8B is a reasoning-focused mixture-of-experts (MoE) model with 0.7B active parameters and 8B total parameters. It is built on Zyphra's MoE++ design, a variant of MoE that improves efficiency and scalability. MoE++ introduces two innovations: Compressed Convolutional Attention (CCA), which reduces computational cost, and the ZAYA1 router, which manages expert selection and improves expert balancing and expressivity. The model's pretraining, midtraining, and supervised fine-tuning were all performed on the full-stack AMD compute, networking, and software platform.

> **"My fundamental message is that AMD works for training at scale."**
>
> Beren Millidge, Co-founder and Chief Scientist, Zyphra

For Co-founder and Chief Scientist Beren Millidge, ZAYA1-8B demonstrates the capability of AMD clusters for large-scale pretraining. He says, "We had to get our stack set up and working on AMD to validate the platform at scale. My fundamental message is that AMD works for training at scale. We have good models coming out of it."

INDUSTRY
AI Research and Product Development

CHALLENGES
Zyphra's previous platform limited model residency per node, increasing parallelism and communication overhead as training scaled for long-context reasoning models

SOLUTION
Zyphra standardized on 1,536 AMD Instinct™ MI300X GPUs, AMD ROCm™ software, and AMD Pensando™ Pollara 400 AI NICs for end-to-end AMD training

RESULTS
Zyphra trained ZAYA1-8B on 20 trillion tokens, achieving a 91.9% AIME'25 benchmark score, 72-hour stable runs, and 16K context without context parallelism, on its way to validating the full AMD stack

AMD TECHNOLOGY AT A GLANCE
AMD Instinct™ MI300X GPUs
AMD Instinct™ MI355X GPUs
AMD ROCm™ software
AMD Pensando™ Pollara 400 AI NICs

Optimizing models for AMD helped Zyphra train larger reasoning models more efficiently while supporting longer context windows.

Using Markovian RSA, its in-house test-time compute method, Zyphra increased ZAYA1-8B performance to 91.9% on the AIME'25 benchmark and 89.6% on the HMMT'25 benchmark, making ZAYA1-8B competitive with substantially larger reasoning models. Zyphra also shaped ZAYA1-8B for efficient AMD training by aligning head dimensions, hidden dimensions, and attention configurations with the GPU architecture.

WHAT 192GB OF MEMORY BRINGS TO TRAINING

Each AMD Instinct MI300X GPU provides Zyphra with 192GB of HBM3 memory, enough to store gradients, optimizer states, and model weights for intensive training workloads. For Zyphra's larger models, this memory capacity reduces reliance on parallelism strategies, allowing larger models and context lengths to be trained efficiently. "The larger memory on AMD Instinct MI300X GPUs lets you fit larger models inside a single accelerator or node without spilling across nodes, which reduces the potential for communication problems."

Because Zyphra focuses on long horizon agentic workloads, context length is a primary concern. "With AMD Instinct MI300X GPUs, we are able to train up to 16K context without needing any context parallelism," Millidge says. "That makes context extension much easier, faster, and more efficient."

> **"The larger memory on AMD Instinct MI300X GPUs lets you fit larger models inside a single or node."**
>
> Beren Millidge, Co-founder and Chief Scientist, Zyphra

Zyphra also had to ensure that its model architecture could handle training at that scale. CCA reduces prefill FLOPs and KV-cache size, and the memory capacity of AMD Instinct MI300X GPUs gives the team the room to take full advantage of those gains. As training advanced from base pretraining into reasoning-focused midtraining and supervised fine-tuning, Zyphra validated both the model design and the AMD stack across increasingly demanding phases.

SCALING ACROSS 1,536 AMD INSTINCT GPUs

Zyphra's cluster runs 192 nodes, each with eight AMD Instinct MI300X GPUs, for a total of 1,536 GPUs. AMD Pensando Pollara 400 AI NICs handle networking across the cluster, giving Zyphra a fully integrated AMD environment for cluster-scale training. All-to-all communication synchronizes gradients and optimizer states across the full system. IBM provided the bare-metal environment and spare nodes to backstop full-cluster runs in the event of hardware issues.

Uninterrupted training time is a key indicator that a cluster is ready for full-scale work. "We trained about 20 trillion tokens on our AMD Instinct MI300X GPU cluster and achieved many uninterrupted 48-hour runs," says Millidge. "We even saw a full weekend, which was about 72 hours. Twenty-four-hour runs now occur regularly, which is excellent."
As Zyphra expanded into the full 192-node cluster, it found that "We have zero problems with the AMD networking stack," Millidge observes. "We get the bandwidth and performance we expected and did not encounter any major networking issues."

> **"The AMD software stack has reached a level of maturity where we can train at scale and trust the GPUs and networking."**
>
> Beren Millidge, Co-founder and Chief Scientist, Zyphra

Stable training at this scale demands full synchronization. "If anything goes wrong, everything goes wrong," Millidge explains. "The AMD software stack has reached a level of maturity where we can train at scale and trust the GPUs and networking to work together. Seeing all of that come together at this scale is really good."

AMD HELPS ZYPHRA PREPARE FOR FULL CLUSTER RUNS

"IBM Cloud and AMD provided several development nodes before the full cluster arrived," Millidge says. "We used those nodes to build and test our training stack on AMD ahead of time, so when the cluster came online, we were not starting from zero."

Reliable AI infrastructure advanced by AMD Instinct™ MI300X GPUs helped Zyphra achieve stable, uninterrupted cluster-scale AI training.

That early work helped Zyphra reach stable training in about two weeks. "It's a testament to AMD, IBM Cloud, and our teams working together to get this running efficiently," continues Millidge.

That early testing focused on the framework and kernel work Zyphra needed for production training. "The ROCm-to-PyTorch conversion was fairly straightforward and does not require too many code changes," Millidge says. "We used the automatic HIP tool to convert kernels from CUDA to HIP, then spent the majority of time optimizing those kernels for efficiency." Zyphra runs its own fork of Megatron-LM and has integrated kernels from AITER AI Tensor Engine for ROCm and Primus-Turbo, with plans to extend that work.

> **"It's a testament to AMD, IBM Cloud, and our teams working together to get this running efficiently."**
> Beren Millidge, Co-founder and Chief Scientist, Zyphra

Since large synchronous training jobs leave little room for even small failures, Zyphra and AMD kept tuning the stack as full-cluster training moved into production. Driver settings, timeouts, network-card behavior, and rare hardware faults all mattered because any one of them could affect an entire run. When one node showed intermittent data corruption, for instance, AMD helped Zyphra isolate and resolve the fault. "Having the fundamental AMD software libraries open-source makes everything much easier and more workable," Millidge says. "We can look at the code directly. Without that visibility, troubleshooting would move much more slowly."

ZYPHRA PLANS THE FUTURE OF SUPERINTELLIGENCE ON AMD

"We're planning to develop even more powerful AI models and to keep building with AMD as we scale toward the frontier," says Millidge. The company is already bringing advanced open-weight models into production through Zyphra Cloud Inference, which uses AMD Instinct™ MI355X GPUs on TensorWave compute infrastructure. Training remains a cornerstone of product development as Zyphra expands its models, advances new architectures, and executes its AI roadmap on AMD technology-based infrastructure.

Powered by AMD, Zyphra Cloud unifies training, inference, and AI agents to help accelerate enterprise AI innovation.

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

ABOUT ZYPHRA

Zyphra is an open superintelligence research and product company based in San Francisco. Zyphra Research pushes the frontier of open intelligence, training multimodal models on heterogeneous compute with a focus on long-term memory, continual learning, and silicon performance. Zyphra Cloud delivers that work at scale with a full-stack AI platform built on AMD, designed for long-horizon agents and serving developers, enterprises, and frontier hyperscalers. For more information visit www.zyphra.com.

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.