November 24, 2025

AMD Powers Frontier AI Training for Zyphra

News Highlights:

- Zyphra ZAYA1 becomes the first large-scale Mixture-of-Experts model trained entirely on AMD Instinct™ MI300X GPUs, AMD Pensando™ networking and ROCm open software.
- ZAYA1-base outperforms Llama-3-8B and OLMoE across multiple benchmarks and rivals the performance of Qwen3-4B and Gemma3-12B.
- Memory capacity of AMD Instinct MI300X helped Zyphra simplify its training capabilities, while achieving 10x faster model save times.

SANTA CLARA, Calif., Nov. 24, 2025 (GLOBE NEWSWIRE) -- AMD (NASDAQ: AMD) announced that Zyphra has achieved a major milestone in large-scale AI model training with the development of ZAYA1, the first large-scale Mixture-of-Experts (MoE) foundation model trained using an AMD GPU and networking platform. Using AMD Instinct™ MI300X GPUs and AMD Pensando™ networking and enabled by the AMD ROCm™ open software stack, the achievement is detailed in a Zyphra technical report published today.

Results from Zyphra show that the model delivers competitive or superior performance to leading open models across reasoning, mathematics, and coding benchmarks—demonstrating the scalability and efficiency of AMD Instinct GPUs for production-scale AI workloads.

“AMD leadership in accelerated computing is empowering innovators like Zyphra to push the boundaries of what’s possible in AI,” said Emad Barsoum, corporate vice president of AI and engineering, Artificial Intelligence Group, AMD. “This milestone showcases the power and flexibility of AMD Instinct GPUs and Pensando networking for training complex, large-scale models.”

“Efficiency has always been a core guiding principle at Zyphra. It shapes how we design model architectures, develop algorithms for training and inference, and choose the hardware with the best price-performance to deliver frontier intelligence to our customers,” said Krithik Puthalath, CEO of Zyphra. “ZAYA1 reflects this philosophy and we are thrilled to be the first company to demonstrate large-scale training on an AMD platform. Our results highlight the power of co-designing model architectures with silicon and systems, and we’re excited to deepen our collaboration with AMD and IBM as we build the next generation of advanced multimodal foundation models.”

Efficient Training at Scale, Powered by AMD Instinct GPUs
The AMD Instinct MI300X GPU’s 192 GB of high-bandwidth memory enabled efficient large-scale training, avoiding costly expert or tensor sharding, which reduced complexity and improving throughput across the full model stack. Zyphra also reported more than 10x faster model save times using AMD optimized distributed I/O, further enhancing training reliability and efficiency. With only a fraction of the active parameters, ZAYA1-Base (8.3B total, 760M

active) matches or exceeds the performance of models such as Qwen3-4B (Alibaba), Gemma3-12B (Google), Llama-3-8B (Meta), and OLMoE.[1]

Building on prior collaborative work, Zyphra worked closely with AMD and IBM to design and deploy a large-scale training cluster powered by AMD Instinct™ GPUs with AMD Pensando™ networking interconnect. The jointly engineered AMD and IBM system, announced earlier this quarter, combines AMD Instinct™ MI300X GPUs with IBM Cloud's high-performance fabric and storage architecture, providing the foundation for ZAYA1's large-scale pretraining.

For further details on the results, read the Zyphra technical report, the Zyphra blog, and the AMD blog, for comprehensive overviews of the ZAYA1 model architecture, training methodology, and the AMD technologies that enabled its development.

Supporting Resources

- Follow AMD on LinkedIn
- Follow AMD on Twitter
- Read more about AMD Instinct GPUs here
- Learn more about how AMD is advancing AI innovation at www.amd.com/aiapplications

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.

Contact:
David Szabados
AMD Communications
+1 408-472-2439
david.szabados@amd.com

Liz Stine
AMD Investor Relations
+1 720-652-3965
liz.stine@amd.com

---

[1] Testing by Zyphra as of November 14, 2025, measuring the aggregate throughput of training iterations across the full Zyphra cluster measured in quadrillion floating point operations per second (PFLOPs). The workload was training a model comprised of a set of subsequent MLPs in BFLOAT16 across the full cluster of (128) compute nodes, each containing (8) AMD Instinct™ MI300X GPUs and (8) Pensando™ Pollara 400 Interconnects running a proprietary training stack created by Zyphra. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations. This benchmark was collected with AMD ROCm 6.4.



Source: Advanced Micro Devices, Inc.