ai& Launches ai& inference: Frontier AI at up to 80% Lower Cost, Built on Heterogeneous Compute

A platform that co-designs across AMD, NVIDIA, and others to push token efficiency beyond single-silicon architecture systems

ai& Launches a Vertically Integrated AI Platform

TOKYO, JAPAN, June 29, 2026 /EINPresswire.com/ -- ai& Launches ai& inference: Frontier AI at up to 80% Lower Cost, Built on Heterogeneous Compute
– A platform that co-designs across AMD, NVIDIA, and others to push token efficiency beyond single-silicon architecture systems –

TOKYO - June 29, 2026 - ai&, a vertically integrated global AI technology company, today launched ai& inference, a high-performance platform that delivers state of the art inference at a fraction of what comparable proprietary inference cost to run today. The platform is built on a heterogeneous compute architecture, combining AMD, NVIDIA, and other silicon architectures under a single optimized inference serving layer. ai& inference is engineered to convert hardware-software co-design into a structural cost advantage that single-architecture providers cannot match.

Competing One Layer Deeper
Inference defines the cost of production AI. As reasoning models and agentic systems push token volume per task ever higher, tokenomics has become critical. Consequently, techniques to push performance have grown in sophistication: speculative decoding, pre-fill/decode disaggregation, aggressive quantization, KV-cache reuse, and model-level routing. While providers have optimized this software layer to a remarkable degree, the industry's reliance on a single source of silicon means every provider ultimately hits the same compute ceiling. Currently,

inference competition is about managing trade-offs within that fixed performance envelope.

ai& competes one layer deeper. It brings that same software discipline to a heterogeneous substrate, integrating AMD, NVIDIA, Tenstorrent, and other architectures natively with the serving stack. ai&'s platform decouples the inference pipeline, executing each computational step on the processor best suited for it. Because ai& owns and operates the hardware end-to-end, it achieves token efficiencies that multiply intense software optimization with hardware co-design. This results in system-level efficiencies that single-architecture providers cannot reach, and a cost structure that compounds favorably every token generated.

In internal benchmarks, this approach delivers substantially higher token efficiency than comparable single-architecture systems on equivalent workloads. This efficiency forms the foundation of the platform's economics, customers see AI quality comparable to leading endpoints, and on agentic and mixed workloads, blended cost up to 80% lower than running every step on a single frontier model. Ultimately, this structural advantage allows ai& to deliver the most competitive pricing against proprietary model providers and amongst inference providers.

Built for Global Enterprise

 Heterogeneous by design. ai& integrates AMD, NVIDIA, Tenstorrent, and other hardware architectures under a unified serving layer. ai& operates the largest AMD-based inference footprint in Japan and largest Tenstorrent deployment globally.
 Agentic cost optimization. A single agentic workflow involves dozens of iterative, multi-turn loops across planning, retrieval, tool use, and verification. ai& Inference handles these intensive cycles by decoupling the inference pipeline and optimizing execution natively for each specific hardware architecture. By maximizing token throughput during these rapid, back-and-forth iterations, the platform dramatically lowers the compounding cost of agentic execution.
 Architectural sovereignty. The serving infrastructure runs entirely on hardware ai& manages directly, eliminating rented cloud infrastructure. For enterprises in regulated industries like financial services, healthcare, and the public sector, inference is served strictly in-region, and in dedicated environments to guarantee full data residency and compliance.
 Regional Low-latency. Serving workloads closer to the end-user removes long-haul round-trip network overhead, delivering the consistent, fast response times required for real-time interactive applications and tight agentic loops.
 Drop-in compatibility. A single configuration change points existing OpenAI- or Anthropic-API-compatible applications directly to ai&'s endpoints. No codebase re-architecting is required to switch.

Owned Infrastructure, Built to Last
ai& has secured more than $2 billion in capital funding to construct multiple 100-megawatt-class AI data centers over the next three years. The platform runs end to end on infrastructure and operates giving customers the performance, economics, and sovereignty that no single-layer

provider can deliver. ai& will continue to expand its heterogenous footprint.

For Customers at Scale
For customers requiring dedicated capacity, custom service-level agreements, on-premise deployment, or specialized workload tuning, ai& offers options that allow customization that can be tailored to unique preferences.

Availability
ai& Inference is available today at console.aiand.com. New users can redeem coupon code UNITABETAI for $50 in free credits.

About ai& Inc.
ai& is a global AI technology company founded on the simple conviction that whoever owns and optimizes the full stack wins. By integrating next-generation data center infrastructure, heterogeneous compute, and frontier model services into a single optimized platform, ai& gives enterprises and developers the performance, economics, and data sovereignty that no single-layer provider can match. Founded in Japan and expanding globally, ai& is building the foundation for the AI-native future.

For more information, visit www.aiand.com, or follow ai& on LinkedIn and X.

# # #

Contact Us
ai&
contact@aiand.com
Visit us on social media:
LinkedIn
YouTube
X