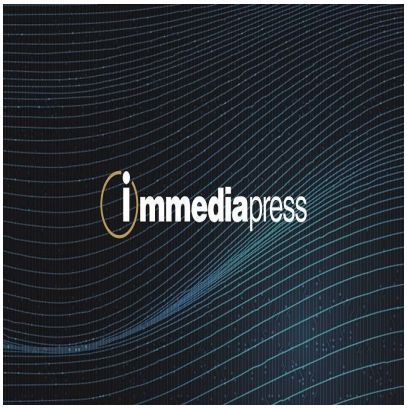

Alibaba Announces Comprehensive Full-Stack AI Upgrade for the Agentic Era

Descrizione

COMUNICATO STAMPA â?? CONTENUTO PROMOZIONALE

HANGZHOU, CHINA â?? Media OutReach Newswire â?? 20 May 2025 â?? Alibaba today announced a comprehensive upgrade of its full AI stackâ??spanning cloud infrastructure and model services, AI chips and foundation models â??to empower customers in building, deploying, and scaling AI agents with greater efficiency, reliability, and performance.

Unveiled at the Alibaba Cloud Summit, Qwen3.7-Max is Alibabaâ??s latest large language model, engineered for advanced agentic coding, complex reasoning, and long-horizon task execution. Qwen3.7-Max will be available soon for developers and enterprises worldwide.

To address surging compute and AI workload demands in the agentic era, Alibaba Cloud has also upgraded its infrastructure and model services. Key launches include the Panjiu AL128 Supernode Server, designed to empower scalable agent inference and large-scale model training, and an optimization update within Alibabaâ??s model service platform that continuously refines model performance.

Additionally, T-Head, Alibabaâ??s semiconductor design subsidiary, introduced the Zhenwu M890, its latest AI training and inference processor, featuring high-capacity memory, robust inter-chip bandwidth, and native FP4 precision support.

Qwen 3.7-Max: A Versatile Foundation Model for the Agent Era

Designed as a robust foundation for AI agents, Qwen 3.7-Max seamlessly handles code generation and debugging, office workflow automation, and complex multi-step tasks requiring hundreds or thousands of actions.

The model delivers exceptional agent capabilities across diverse domains. As a frontier-level coding assistant, it supports coding tasks from rapid frontend prototyping to complex, multi-file software engineering. To enhance office work productivity, it reliably orchestrates multi-agent workflows to tackle sophisticated operations. Notably, Qwen 3.7-Max can autonomously execute long-horizon agentic tasksâ??sustaining continuous operation for up to 35 hours and managing over 1,000 tool calls without performance degradation.

Deeply optimized for leading agent frameworks including OpenClaw, Hermes Agent, Claude Code, Qwen Paw and Qoder, it serves as a reliable backbone for different agent systems. The model achieves top-tier results across major benchmarks in coding, general-purpose agents, general capabilities and multilingualism, making it competitive with leading frontier models. It will be soon accessible through Alibabaâ??s model service platform Model Studio for global developers.

Next-Generation Intelligent Computing and Enhanced Model Services

To empower scalable AI Agent inference and large-scale model training, Alibaba Cloud has launched the Panjiu AL128 Supernode Server, powered by the Zhenwu M890 AI processor and ICN Switch 1.0 networking chip. By tightly integrating 128 AI accelerators within a single rack, the system delivers single-rack bandwidth at the petabyte-per second (PB/s) scale, dramatically improving the handling of large-scale concurrent requests from agents.

The Panjiu AL128 is now available on Model Studio for the China market (or â??Bailianâ?•), enabling Chinese enterprises to efficiently address training and inference demands across sectors.

To optimize performance, Bailian has introduced Agentic RL, a reinforcement learning mechanism powered by agent execution feedback, to drive continuous model iteration. Bailian also features built-in safety governance capabilities, ensuring that autonomously operating agents always remain within defined boundaries.

T-Headâ??s Latest Chips and Software Stack for AI Training and Inferencing

T-Headâ??s latest AI accelerator, the Zhenwu M890, delivers three times the performance of its predecessor Zhenwu 810E. Zhenwu M890 features 144 gigabytes (GB) of GPU memory and 800 GB per second of inter-chip bandwidth. The chip natively supports multiple data precision formats, ranging from FP32 (32-bit floating-point) down to FP4 (4-bit floating-point), supporting both high-precision model training and ultra-low-precision model inference. These capabilities make it exceptionally well-suited for complex agentic AI workloads, which demand extensive working memory for context retention, high-speed communication for multi-agent coordination, and low-precision computing to maintain rapid execution while reducing cost. The chip is built on T-Headâ??s proprietary parallel computing architecture and utilizes its custom ICN (Inter-Chip Network) interconnect protocol.

Alongside the accelerator, T-Head unveiled the ICN Switch 1.0, a dedicated switching chip designed to create high-bandwidth, low-latency scale-up networks for compute clusters. It delivers up to 25.6 Tbps

of aggregate bandwidth and achieves extreme low latencyand congestion-free communication. By pairing the Zhenwu M890 with the ICN Switch 1.0 chip, it enables full-bandwidth interconnection across 64 accelerators, significantly boosting the computational efficiency and stability of large-scale intelligent computing. T-Head also unveiled its proprietary software stack, T-Head SAILâ?¢, to unleash the full computational potential for its chips.

T-Head has achieved widespread industrial adoption of its proprietary AI chips, with over 560,000 Zhenwu units delivered to date. More than 400 external customers across 20 industries, including leading automakers and financial services companies, have deployed the chips to power intelligent operations.

About Alibaba Group

Alibaba Group is a global technology company focused on e-commerce and cloud computing. We enable merchants, brands and retailers to market, sell and engage with consumers by providing digital and logistics infrastructure, efficiency tools and vast marketing reach. We empower enterprises with our leading cloud infrastructure, services and work collaboration capabilities to facilitate their digital transformation and grow their businesses.

Contatti:
Immediapress
comunicati@immediapress.it

The media contact for usage if required:Crystal Liu

Crystal.liu@alibaba-inc.com

COMUNICATO STAMPA â?? CONTENUTO PROMOZIONALE

ResponsabilitÃ editoriale di Immediapress

â??

immediapress

Categoria

1. Comunicati

Tag

1. ImmediaPress

Data di creazione
Maggio 20, 2026
Autore
redazione