FOR IMMEDIATE RELEASE

Alibaba Cloud Unveils Strategic Roadmaps for the Next Generation AI Innovations

Full-stack offerings introduced from AI models to agent development and application platforms and upgraded infrastructure during Apsara Conference 2025

Hangzhou, China, September 24, 2025 – Alibaba Cloud, the digital technology and intelligence backbone of Alibaba Group, today unveiled its latest full-stack AI innovations at Apsara Conference 2025, its annual flagship technology conference. The announcement spans from next-generation large language models from the Qwen3 family, the upcoming Wan 2.5 visual-generation models, enhanced platforms for agent development and application, to major upgrades of its AI infrastructure, reinforcing the company's global leading position at the forefront of the new AI era.

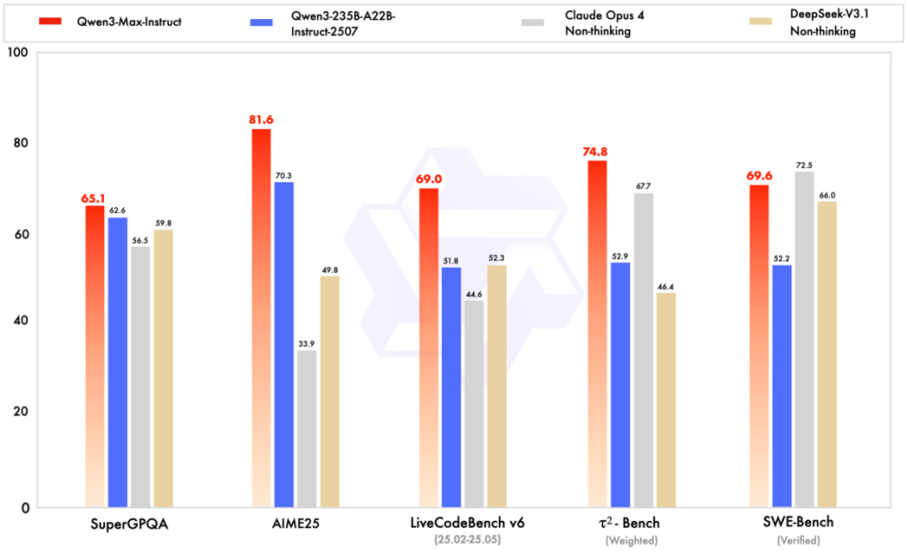
"In the future, large AI models will be deeply integrated into a wide range of devices, functioning like operating systems — equipped with persistent memory, seamless cloud-edge coordination, and the ability to continuously evolve. **We remain committed to open-sourcing Qwen and shaping it into the 'operating system of the AI era,**' empowering developers around the world to build transformative AI applications," said **Eddie Wu, Chairman and CEO of Alibaba Cloud Intelligence.**

"Simultaneously, **Alibaba Cloud is strategically positioned as a full-stack AI service provider, dedicated to delivering robust computing with maximized efficiency for training and deploying large AI models on the cloud**. To underscore our long-term commitment to advancing AI, we will progress with our RMB 380 billion investment plan in AI and cloud infrastructure over the next three years," Wu added.

Since the launch of the first generation of Qwen in 2023, **Alibaba has open-sourced over 300 AI models** built on its two foundation models: the large language model Qwen and the visual generation model Wan. **With over 600 million downloads and 170,000+ derivative models** created, Alibaba's AI models have become one of the most widely adopted open-source AI series globally. Notably, **over 1 million corporates and individuals** have used Qwen on Model Studio, Alibaba's AI development platform.

Unveiling Qwen3-Max: The most powerful LLMs from Alibaba to date

Alibaba officially released **Qwen3-Max**, its largest LLM model with over 1 trillion parameters. With Instruct (non-thinking) and Thinking modes, the model achieves impressive performance across a wide range of benchmarks especially in code generation and agentic capabilities. For the instruct mode, it scores 69.6 in **SWE-Bench**, an authoritative benchmark for evaluating LLMs on real-world software issues, on par with some leading closed-source models. It also records remarkable performance on **Tau2-Bench**, a benchmark that evaluates conversational agents, showing exceptional proficiency in tool use, a foundational capability for building intelligent, action-oriented agents.



Qwen3-Max achieves impressive performance across benchmarks

A series of Qwen3 models that cover visual language and multimodal processing were also unveiled at the conference.

- **Qwen3-VL:** The most capable vision-language model in the Qwen family to date. Its Mixture-of-Experts (MoE) architecture enables flexible deployment from edge devices to high-performance cloud environments. Functioning as a **visual agent**, Qwen3-VL is capable of operating on both computer and mobile interfaces; It pioneers **visual programming** by generating code directly from images or videos, effectively turning visual designs into functional applications. Its **spatial understanding** capability support 3D grounding with enhanced perception of direction and distance, laying critical groundwork for embodied AI and real-world spatial navigation. **Qwen3-VL-235B-A22B** is available in both Instruct (non-thinking) and Thinking versions, achieving remarkable performance across leading visual perception and multimodal reasoning benchmarks.

- **Qwen3-Omni:** a natively end-to-end, multilingual omni-model capable of processing text, images, audio, and video inputs, while delivering real-time, streaming response in both text and natural speech. Powered by a novel Thinker–Talker MoE architecture and pre-trained on 20 million hours of audio data, Qwen3-Omni delivers exceptional performance in understanding audio input (up to 30 minutes) and video-based conversation, all without compromising its strong capabilities in text and image processing. It also achieves real-time multimodal interaction, with ultra-low latency – making it an ideal solution for intuitive, hands-free interaction in intelligent cockpits, smart glasses and mobile phones. **Qwen3-Omni-30B-A3B** is now open sourced on Hugging Face and Alibaba Cloud's ModelScope community. Users can also access **Qwen3-Omni-Flash** on Qwen Chat, a web application that allows users to experience different Qwen models.

Additionally, **Qwen3-Coder and Qwen3-Image-Edit have received a major upgrade**. The new Qwen3-Coder achieves faster inference speed and enhanced code safety, while Qwen3-Image-Edit has been updated to support multi-image editing with significantly improved visual consistency.

Alibaba also unveiled **Fun**, a family of speech LLMs equipped with advanced multilingual speech recognition and synthesis capabilities. The series includes **Fun-ASR,** an end-to-end automatic

speech recognition (ASR) model optimized for real-world enterprise deployment, and **Fun-CosyVoice**, a high-quality, expressive speech synthesis model designed to generate natural-sounding spoken output in multiple languages.

Wan2.5 Preview: Elevates Multimedia Content Creation

At the same conference, Alibaba also **previewed four Wan2.5 models**, including its latest video generation models, an image generation model and an image editing model. The video generation models natively support high-fidelity audio generation for the video, doubling the duration from 5 to 10 seconds, enabling more complete and coherent narratives with enhanced visual quality. The models feature a natively integrated multi-modal architecture, which is trained jointly on text, audio, and visual data. This allows for aligned multi-modal generation, ensuring synchronized audio and visual content, and enhanced instruction understanding to closely follow user prompts.

New Development Framework for Enhanced Agent Deployment

For improved efficiency of implementing AI agents at scale, a development framework is now added to Model Studio, Alibaba Cloud's AI development platform. The new framework features **Model Studio-ADK** (agent development kits), **a high-code development framework for enterprise professionals** that translates intricate business needs into executable agent logic to enables the rapid development of sophisticated AI agents with autonomous decision-making, dynamic reflection, and iterative task execution capabilities. Remarkably, users can create a DeepResearch or Agentic RAG (Retrieval-Augmented Generation) project within an hour using this robust toolkit. Model Studio has also upgraded its low-code development platform Model Studio-ADP (Agent Development Platform), enabling users with limited programming backgrounds to easily create lightweight AI agents.

Addressing key enterprise challenges such as multi-source data processing, resource constraints, and cross-environment deployment, **Model Studio Agent introduces a range of enterprise-grade features**. These include seamless connectivity via Model Context Protocol (MCP), RAG multi-modal fusion, dynamic inference scheduling, and sandbox service, allowing enterprises to accelerate the adoption of AI agents.

Currently, users can access over 200 industry leading models via Model Studio, including Alibaba's self-developed Qwen and Wan models. **More than 800,000 agents have been created on Model Studio,** supporting diverse scenarios ranging from content creation and intelligent marketing to smart home management and production optimization. **Over the past 12 months, number of model calls via Model Studio have increased by 15 times**, reflecting the growing demand for robust and scalable AI solutions.

Novel AI Platforms to Support Enterprises and Creators

Following its debut in July, Alibaba Cloud has rolled out major upgrades to **AgentBay**, a multimodal cloud-based operating environment and expert agent platform for enterprises, developers, and AI partners. The new features—Self-Evolving Engine, custom container images and builtin safety and compliance controls—help transit agents from simple, single model helpers to composite, human-like, multimodal workers that can complete tasks end-to-end.

To meet rising enterprise demand for AI-driven growth, Alibaba Cloud also launched **Lingyang AgentOne**, **a one-stop enterprise AI application platform** that enables organizations to move from reactive response to proactive intelligence. Powered by Alibaba's Qwen models and deeply integrated with the Alibaba ecosystem, Lingyang AgentOne offers an end-to-end agent development

workspace to connect with existing systems and accelerate time-to-value. Through scenario-based solutions across marketing, analytics, customer service, and operations, Lingyang AgentOne links the full pre-sales, sales, and post-sales value chain to deliver measurable, production-ready outcomes for industries such as home improvement and e-commerce.

Additionally, Alibaba's consumer-facing AI application platform **Quark launched Zaodian**, **a one-stop AI image and video creation platform** that integrates industry leading AI models such as Alibaba's flagship video generation model Wan to deliver a professional, efficient experience for creators. Apart from the text-to-video and image-to-video functions supported by Wan, Zaodian also offers AI image generation and editing functions with leading model choices. Creators can experience the platform service at website zaodian.quark.cn or via the “AI Image” entry on Quark desktop version.

Next-Generation AI Infrastructure for Agentic AI

At the conference, the cloud pioneer has also unveiled a comprehensive suite of innovative infrastructure upgrades specifically designed to support the emerging agentic AI landscape.

- **Storage**: Alibaba Cloud enhanced its Object Storage Service (OSS) with “Vector Bucket,” an AI-powered feature enabling cost-efficient, large-scale vector data storage and retrieval — optimized for RAG and AI apps. It unifies raw and vector data management in OSS, accessible via standard APIs, simplifying scalable RAG platform development and multimedia asset organization. It helps lower the cost of AI development by letting businesses manage both raw and vector data in one place — reducing complexity and accelerating RAG application deployment.

- **Networking**: Alibaba Cloud unveiled its latest architecture for high performance network—HPN8.0, a network specially designed for AI models. This innovation enables seamless model training, inference, and reinforcement learning (RL) across mixed computational workloads, while supporting ultra-large-scale deployments. The architecture delivers 800 Gbps network throughput, doubling the capacity in previous generation.

- **Security**: Another key update is the addition of an AI-driven agentic function to its Cloud Threat Detection Response (CTDR) solution. This cloud-native security enhancement boosts detection, analysis, and response capabilities, providing a more intelligent and proactive approach to combating security threats. Five AI agents, powered by Qwen, will automate security operations—from alert assessment to execution—with intelligent analysis, event correlation, and actionable reporting for end-to-end threat management. The new function has effectively increased the automated incident investigation success rate from 59% to 74%, while handling 70% automated response actions without human intervention.

- **Container:** Alibaba Cloud has upgraded its Container Compute Services (ACS) to enhance its auto-scaling capabilities through optimized scheduling and container image cache acceleration technologies. This enables elasticity, supporting the scaling of up to 15,000 pods per minute to handle massive, highly concurrent agent requests. Besides, the ACS container sandbox technology provides strong isolation by separating user space from the runtime environment, preventing vulnerabilities or data leaks in one agent from affecting others.

- **Database**: Alibaba Cloud has upgrades its PolarDB database, optimizing for combined data and AI workloads. The upgrade has introduced a hardware innovation powered by Compute Express Link (CXL) technology, a highly efficient compute-memory interconnect to reduces latency by 72.3%, boosting memory scalability by 16x and laying a solid foundation for data and AI workload. The upgraded PolarDB also introduced a new Lakebase architecture with hybrid storage include lake, operational database and metadata for storing popular open-data formats including Lance, Iceberg and Apache Hudi and lowering storage cost, enabling efficient multimodal data storage and management.

- **Platform for AI (PAI)**: Alibaba Cloud's PAI introduced synergistic optimizations to advance large model development into the agentic AI era. Its novel MoE training acceleration improves Qwen series training by over 300%, while the upgraded DiT training engine reduces Wan series' single-sample training time by 28.1%. Enhanced inference delivers 71% higher TPS, 70.6% lower TPOT latency, and 97.6% faster infrastructure scaling.

###

About Alibaba Cloud
Established in 2009, Alibaba Cloud (www.alibabacloud.com) is the digital technology and intelligence backbone of Alibaba Group. It offers a complete suite of cloud services to customers worldwide, including elastic computing, database, storage, network virtualization services, large-scale computing, security, big data analytics, machine learning and artificial intelligence (AI) services. Alibaba has been named the leading IaaS provider in Asia Pacific by revenue in U.S. dollars since 2018, according to Gartner. It has also maintained its position as one of the world's leading public cloud IaaS service providers since 2018, according to IDC.

Media Contacts

Crystal Liu
Crystal.liu@alibaba-inc.com
+8618578497650/+852 60192703

Luica Mak
Luica@alibaba-inc.com
+44 7905471332