FOR IMMEDIATE RELEASE

Alibaba Unveils Qwen4-Max: New AI Teammate for Work, Creativity, and Discovery

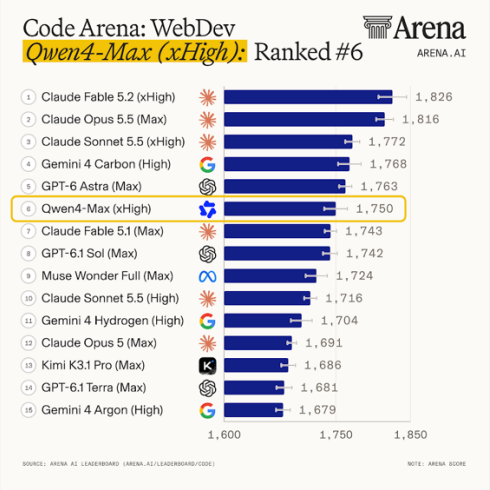
Hangzhou, China, November 16, 2026 - Alibaba has released Qwen4-Max, the most capable model in the Qwen family to date. Featuring 2.4 trillion total parameters, native vision capabilities, and a one-million-token context window, Qwen4-Max excels at complex tasks across work, creativity, and research.

Qwen4-Max achieves top rankings across leading benchmarks, particularly in coding, agentic tasks and visual reasoning. It ranked 6th on Code Arena, a real-world coding leaderboard, and posted strong scores across specialized benchmarks for agent workflows, workplace productivity, and visual reasoning, including Agents' Last Exam, Vals Index, JobBench, and Chartography.

According to Work360—a proprietary benchmark evaluating performance on challenging industry tasks across 721 specialized fields—Qwen4-Max outperforms its predecessor, Qwen3.8-Max, overall while using 35% fewer tokens at High effort reasoning mode and reducing the average cost per task by around 60%.

These gains are driven by the new Qwen4 architecture, which upgrades the sparse Mixture-of-Experts (MoE) design across attention, residual connections, and embeddings. **Built as an MoE model with 2.4 trillion total parameters, Qwen4-Max activates only 71 billion per token—25% fewer activated parameters compared to Qwen3.8-Max at the same 2.4T scale—delivering stronger capabilities at a lower inference cost.**

Qwen4-Max is available now on QwenCloud via the Token Plan and API. Users can also access the model across Alibaba platforms, including QwenWork and QwenStudio. Input (Implicit Cache Hit) is priced at $0.075 per 1M tokens. **Open-weight, smaller-scale models from the Qwen4 series will soon be released to developers worldwide.**



Qwen4-Max ranked 6th on Code Arena



Qwen4-Max outperforms Qwen3.8-Max overall on Work360

Improving Resource Efficiency and Enhancing Productivity

Qwen4-Max features advanced Cloud Resource Scheduling capabilities designed for deployment in cloud infrastructure. Operating in a closed loop, the model autonomously optimizes its own scheduling workflows to cut costs and boost resource efficiency. In benchmark testing, this reduced data-annotation model call costs by an average of 9%, and accelerated offline CPU rescheduling speeds by more than 14 times.

Beyond cloud infrastructure, **Qwen4-Max supports professional workflow and creative production** such as 3D modeling, game development, and agentic video creation. In **automotive manufacturing**, Qwen4-Max delivers end-to-end capabilities across production-line planning, part design, and machining verification. During a showcase, the model cut manual workpiece programming time by 95%—from roughly 475 minutes to just 24 minutes—by using feedback-driven iteration to catch overcuts

and insufficient material removal before machining begins. Qwen4-Max also accelerates forward-deployed engineering (FDE) across industries such as finance, law, healthcare.

Powering Long-Running Agents for Scientific Discovery

Built for the agentic AI era, Qwen4-Max powers sustained, around-the-clock execution. In the science field, Qwen4-Max's long-horizon planning helps bridging digital and physical scientific discovery. To tackle next-generation chip cooling challenges, **Qwen4 completed a continuous 48-hour agent run through four iterations of a closed-loop workflow**—combining molecular design, synthesis and testing, and active learning—**to discover an experimentally validated coolant molecule.** Similarly, **in an Alzheimer's research experiment, Qwen4-Max demonstrated end-to-end capabilities** across target research, molecular design, and computational evaluation. Across the workflow, it autonomously executed 454 tool calls and 3,239 metric evaluations to design and optimize 24 candidate molecules.

Safety Built by Design

Alibaba has embedded model governance throughout the entire lifecycle. **Software engineering training included code security checks alongside functionality tests**, significantly reducing vulnerabilities like injection, memory safety, and broken access control. **For cybersecurity, Alibaba filtered training data by task intent, retaining defensive material for vulnerability detection and remediation while removing offensive attack content. Additionally, dedicated safety datasets and post-training enhancements further strengthen protections against chemical, biological, radiological, and nuclear (CBRN) risks.**

###

About Alibaba Group
Alibaba Group is a global technology company focused on AI + Cloud and commerce. We empower consumers and enterprises with our full-stack AI capabilities and services, from applications to compute infrastructure. Our AI technology based on the Qwen family of large language and multimodal models powers the intelligence behind our services across enterprise solutions and consumer platforms. Our commerce business puts consumers first and provides the technology and marketing reach to help merchants, brands, retailers and small businesses to engage with customers and operate efficiently.

Media Contact
Crystal Liu
Crystal.liu@alibaba-inc.com

Luica Mak
luica.mak@alibaba-inc.com