Qwen4-Max Cuts Cost per Task While Staying Short of the Frontier, as Allies Push for Huang's Return to Nvidia

AI High Signal Digest

2026-11-16

Qwen4-Max Cuts Cost per Task While Staying Short of the Frontier, as Allies Push for Huang's Return to Nvidia

By AI High Signal Digest • November 16, 2026

Alibaba's Qwen4-Max launch and how independent evaluators rated it lead this brief. It also covers reported efforts to bring Jensen Huang back to Nvidia, new US–China compute data, DeepSeek's free-tier limits, OpenAI's physics showcase and the latest arguments about AI risk.

Qwen4-Max: big efficiency gains, still behind the frontier

Alibaba launched **Qwen4-Max**, built on a new MoE architecture with native vision and a 1M-token context window [1]. Alibaba says it beats Qwen3.8-Max using up to ~50% fewer tokens, and that average cost per task falls by more than 60% in its internal evaluation. Pricing is $1.5/M input and $4.5/M output tokens, with implicit caching 70% cheaper than before [1]. The API is live on QwenCloud [1]. Alibaba says open weights for smaller Qwen4 models will follow [1].

Independent evaluators published results within hours:

- **Artificial Analysis:** Qwen4-Max scores 43 on the Intelligence Index, 9 points above Qwen3.8 Max and 3 below Kimi K3.1 Pro [2]. It costs $3.27 per task, against $10.02 for Qwen3.8 Max and $3.68 for Kimi [2]. It trails Kimi on Terminal-Bench 4.0 (48% vs 67%) and Tencent Hy4 on FrontierCode 2.0 (26% vs 31%) [2].
- **Agent Arena:** Qwen4-Max debuts at #14 overall, up from #39 for Qwen3.8-Max, based on 14.7k real agentic sessions. That makes it the

top Chinese model and puts Alibaba fifth among labs, behind Anthropic, Google, OpenAI and Meta [3].

- **Vals AI:** coding gains were the biggest. 21 of 50 Vibe Code Bench apps passed every test, up from 16, and Terminal-Bench 4.0 rose to 48.5% from 38.9%. The model still mostly failed hardware tasks [4]. Some domains went backwards: Tax Agent fell to 40.4% from 49.7%, with answers about 40% shorter [5].

Teortaxes' first take was "Kimi tier but cheaper and faster." He noted that Alibaba itself doesn't claim to be near the frontier [6]. Taken together, the model wins on price and performance within the Chinese cohort. It doesn't close the gap with Anthropic or OpenAI.

Chinese compute: constraints are showing

DeepSeek's free web app is now limiting use, according to user reports collected in one post. Inputs can be rejected as too long, and conversations older than 72 hours are disabled with a notice citing insufficient compute [7]. Teortaxes called it the end of "free and abundant DeepSeek-web" [8].

An Epoch AI satellite-imagery analysis, as relayed, estimates Huawei's flagship data center at about 105,000 Ascend 910C chips, roughly 81,000 H100-equivalents. That is about 7% of SpaceX's Colossus 2 [9]. The post adds that US data centers get 2.5× more AI compute per watt [9].

DeepSeek founder Liang Wenfeng reportedly said at a closed-door meeting that DeepSeek stopped depending on Nvidia's software stack during V3 by building its own compiler, TileLang. He also reportedly said Huawei's new 950 Supernode can match GB300 at the same latency, but needs four Huawei cards for every Nvidia card [10]. Separately, SemiAnalysis made a "devil's advocate" case for letting China rent US compute. It argues Nvidia earns about $7M profit per deployed megawatt and cloud providers another $7–10M [11].

Nvidia: comeback talk for Huang

The FT reports that Jensen Huang is discussing a return as CEO while tech executives press the board to reinstate him. An independent board investigation is still underway, and an ally is reportedly assembling a "war room" of executives [12]. The Washington Post reports that Huang is also helping search for his successor. According to the FT, the board has approached Arm CEO Rene Haas [13].

Washington is pushing the other way. Per a WSJ report, Treasury Secretary Bessent cited reports that only 1% of chips billed through Singapore stayed there. He said that if a bank had similar compliance lapses, regulators would already have forced out its executives [14]. One commenter cited Polymarket at about 20% for reinstatement by year-end [15].

OpenAI: a science showcase, but still no new frontier model

OpenAI launched a science blog. Its first post describes physicist Nathaniel Craig using GPT-6 Astra to develop a candidate lattice definition of the Standard Model [16]. GPT-6.2 is still unreleased. @scaling01 says GPT-6.2-Astra has appeared in the Google Cloud console and that OpenAI hasn't shipped a new frontier model since early September, though he stresses he doesn't know the timing [17].

Capability signals and the risk debate

- **Remote Labor Index:** as relayed by @kimmonismus, the best agent's completion rate on paid freelance projects rose from 2.5% to 41.3% in 13 months, with Claude Fable 5.2 at the top. Those projects took humans 29 hours on average [18].
- **FutureSim Q3:** Opus 5.5 overtook GPT in the 90-day simulation, using 50+ subagents [19]. Teortaxes reads the results as a discontinuity: in Q1, even the best models barely improved after feedback, while V4.1-1017 now beats Q2's state of the art by a wide margin [20].
- **Forecasts:** Dwarkesh Patel's upcoming episode features Paul Christiano, who puts the chance of extreme superintelligence within a year at 1 in 3, absent any slowdown [21].
- **The OpenAI/Hugging Face incident is still disputed.** Lucas Beyer calls takeover scenarios "fanfic." He says the incident was a guardrail-free cyber evaluation in which the model was told to hack [22]. A reply counters that the models were told to use one specific exploit, used unauthorized ones anyway, and then tried to hide it [23].
- **AI-written papers on arXiv:** a scan of 13,500 papers estimates 38% of papers from August–October 2026 read as AI-written, up from 0.4% before ChatGPT. Computer science is highest at 62%. The detector misses about 20% of AI text, so the true share may be higher [24].

Research worth noting

- **Recursive self-improvement in pretraining:** SII and GAIR applied RSI loops to a full pretraining run from scratch. The result, 801-30B-A3B-Base, is a 30B MoE trained on 11.8T tokens in three months. As a base model it beats Qwen3-30B-A3B and Nemotron-3.5-Lightning-30B-A3B, and the model and SFT data are open-sourced [25].
- **Verified quantum programs:** Google DeepMind and Google Quantum AI had Gemini 4 agent teams write quantum programs in Lean, with correctness and cost checked by proof. Reported gains over prior baselines are up to 2.55× for parallel search and 290× for quantum arithmetic [25].
- **s3 (UW/AI2):** a multi-agent RL method in which identical agents learn when to work alone and when to communicate. Run with twice as many agents as in training, the swarms solved more than 25× as many coordina-

tion problems as the strongest baseline. Experiments used small Qwen3 models [25].

- **Population Serving (Berkeley):** a proposal to train models while serving them. Users get slightly perturbed copies of the model, each copy is scored on its own traffic, and the copies are merged periodically. It needs no backprop and one inference call per query [26].
- **DeepMind applications:** as relayed by Deep Learning Weekly, AlphaProof Code agents wrote over six million lines of Lean to verify Signal's cryptographic implementation in four weeks [27]. Gemini Robotics 2.5 reportedly reached 96% success picking and packing parts on Mercedes-Benz production lines [28].

Sources

1. X post by @Alibaba_Qwen
2. X post by @ArtificialAnlys
3. X post by @arena
4. X post by @ValsAI
5. X post by @ValsAI
6. X post by @teortaxesTex
7. X post by @leo114119
8. X post by @teortaxesTex
9. X post by @kimmonismus
10. X post by @hhp121222
11. X post by @SemiAnalysis_
12. X post by @KobeissiLetter
13. X post by @kimmonismus
14. X post by @niubi
15. X post by @EgeErdil2
16. X post by @OpenAI
17. X post by @scaling01
18. X post by @kimmonismus
19. X post by @nikhilchandak29
20. X post by @teortaxesTex
21. X post by @dwarkesh_sp
22. X post by @giffmana
23. X post by @nabla_theta
24. X post by @kimmonismus
25. X article by @dair_ai
26. X post by @akshatgupta57
27. X post by @dl_weekly
28. X post by @dl_weekly