SDS PODCAST

EPISODE 1018: ALIBABA'S QWEN3.8-MAX: OPEN-WEIGHT MODEL SURPASSES MOST AMERICAN FRONTIER LABS

Jon Krohn: 00:00 This is episode number 1018 on Alibaba's Quen 3.8 Max. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. Today's episode is on Quen 3.8 Max, the enormous new flagship model from the Chinese tech giant Alibaba, a model that if Alibaba keeps a promise, it made it launch, is now the largest open weight AI model release in history. If that setup sounds familiar, well, it should. Three weeks ago in episode number 1012, I covered Kimi K3, the nearly three trillion parameter model from Beijing-based moonshot AI that rattled investors, kicked off a pricing skirmish among the big American AI labs, and reignited the open source AI debate in Washington. Well, here's a fun fact. Moonshot is backed by Alibaba, and now Alibaba itself has stepped into the ring with its own frontier scale release. The timing throughout has been pointed. Alibaba previewed Quen 3.8 Max at the World AI Conference in Shanghai on July 19th, two days after Moonshot released K3, and then shipped the full model on August 3rd.

01:11 Alibaba's Hong Kong listed shares rose 7% on launch day, which tells you how much investor sentiment now rides on these releases. And this is all part of a much bigger bet. Alibaba has committed more than 380 billion one, that's about 50 billion US dollars to cloud and AI infrastructure over three years. It is dwarfed in comparison to what some of the big tech companies are doing in the US, but it is nevertheless a big sum on AI infrastructure for sure. So what is the Quenn 3.8 Max model and why is it such a big deal? Well, it's a 2.4 trillion parameter model, the most capable in the Quen family to date. And like KIMI K3, it's a mixture of experts architecture. This means that of those 2.4 trillion total parameters, only a small proportion, about 95 billion are active for any given token. So the headline number describes the model's total capacity rather than the compute and the cost burned on every request.

02:14 The model accepts texts, images, and video as input, returns text as output, and supports a one million token context window, enough to hold roughly three quarters of a million words in a single query matching both K3 and the American flagships on that context window dimension. It's built on the architectural foundation of QUEM 3.5 scaled way up. And one practical detail I appreciate where K3's reasoning mode is always on and locked to maximum effort, a pricing gotcha, I flagged back in episode number 1012, QUEN 3.8 Max lets you select low, medium, or extra high reasoning effort so you're not paying for a lengthy thinking trace when all you need is a quick lookup. Now, how good is it? Alibaba's own framing is that QUEN 3.8 Max is second only to Anthropics Claude Fable five or Methos five, depending on what you have access to. And while that's a vendor self-assessment, the independent signals so far land in a similar neighborhood, which is jaw dropping and for Western foreign policy hawks, probably concerning.

03:17 Let me repeat that. An open weight model may lag behind only one model, Anthropics Fable five, which is eye-wateringly expensive to use and was presumably even more eye wateringly expensive to create. On the popular crowdsourced arena platform, Quen 3.8 Max immediately became the highest ranking Chinese model for text tasks. That's tough to say, text tasks, though it still trails several anthropic offerings, including Fable five. And on vision tasks, it ranked second globally behind only a Fable five variant. On the Artificial Analysis Intelligence Index, which is another widely reputed index for tracking AI capabilities, Quinn 3.8 Max scores 56. For reference, Kimi K3 scored 57. So by that measure, these two Chinese giants are in a statistical dead heat near the top of the open weight pack. On specific benchmarks, there are wins to point to as well. Alibaba's published table has Quen 3.8 Max at 86.6 on Terminal Bench, a benchmark of agentic command line tasks ahead of both Claude Opus 4.8 and QuadFable five at 84.6, though behind

OpenAI's GPT 5.6 Soul working in its maximum effort mode at 88.8.

04:35 That sounds expensive to run. With this model, Quen 3.8 Max, where Alibaba is pushing the hardest, however, is on long horizon autonomy. In one internal demonstration, the model spent more than 10 days autonomously coding a self-evolving software harness from scratch, incorporating user feedback, running its own tests, and iterating through code, previews, and logs without human help. In another demo, it reproduced a machine learning research paper from Xero running 33 rounds of GPU training over roughly 125 hours, writing 7,600 lines of code, and then devising 18 improvement ideas that ended up outperforming the original paper's method. Entered into a live online contest against 526 human teams. It beat 87% of the field. Those are vendor demonstrations, so hold them loosely until third parties reproduce similar kinds of results, but the direction is unmistakable. Multi-day unattended agentic work is the battleground these frontier labs are fighting on now.

05:37 And yeah, then there's the pricing story. Let me dig into that a bit more now. Quinn 3.8 Max costs $2 per million input tokens, $6 per million output tokens, and 25 cents per million cashed input tokens. Recall from episode number 1012 that K3 charges $3 in and $15 out. So Quenn 3.8 Max undercuts its own domestic rival by quite a bit, including by more than half on output tokens. Against the American labs, the gap is much wider still. That $8 combined rate adding up input and output is less than a third of Claude Opus's combined pricing forget fable and under a quarter of GPT 5.6 soles. And because cached input is eight times cheaper than fresh input, agentic and retrieval heavy workloads with stable system prompts collapsed toward that 25 cent per million token floor. Wow. The price war I described three weeks ago has not cooled.

06:33 It has escalated. Now for the asterisk, and it's the same asterisk Kimi K3 carried at launch. At the time of me recording what I'm saying to you right now, which is a few days before this episode is released, the open weights for the model are promised, but not yet shipped. Alibaba has committed to publishing weights for Quen 3.8 Max, the first max class Quen model ever to go open during the week. I am recording this on both Hugging Face and Modelscope alongside a smaller Quen 3.827B that will fit on a single high-end GPU for those of us without a data center handy. Hopefully by the time of publication, you can access those model's weights as you are expected to be able to do. Recent Quen open releases shipped under the permissive Apache 2.0 license, which sets an encouraging precedent and hopefully the Quen 3.8 max weights get the same permissive treatment.

07:28 Now, all of this brings me to a question I get asked a lot and want to spend the last stretch of this episode on. Are these Chinese models safe to use? Well, the risk depends far less on the model and far more on how your data reach it. At one end of the spectrum are the consumer services. So chatbot apps and workplace platforms like Alibaba's new Quen Work. Whatever you type into those is collected by the provider and stored on infrastructure governed by Chinese law. So I'd advise against putting anything sensitive, proprietary, or personal into those. One step down in risk is the hosted API on Alibaba Cloud. You get contractual terms and enterprise controls, but your data is still transit a Chinese company server. So check your compliance obligations before writing customer information through it. And again, I don't know if I would put sensitive proprietary or personal info through those.

08:18 Safer again is using a Western cloud provider like Lightning AI where I hold the fellowship or use a model gateway that hosts the open weights on infrastructure in your own jurisdiction. Your prompts never touch Chinese servers at all in that circumstance. But at the safest end of all, it's the option that Openweights uniquely unlock downloading the model and running it entirely on your own hardware. Weights are inert files of numbers. They can't phone home and you can run them fully air gapped if you wish with no per token fees and no data leaving your walls. Even then, two caveats apply. First, self-hosting doesn't change what's inside the model. Its outputs reflect its training, including alignment with Chinese content regulations on politically sensitive topics. So evaluate it on your own use cases before trusting it. Second, practice basic supply chain hygiene. Download from the official repository, prefer the safe tensors format and verify check sums.

09:16 And note that some governments and regulated industries restrict Chinese origin models regardless of where they're deployed. So check the rules that apply to you. Zooming out, the pattern from episode number 1012 three weeks ago on Kimmy K3 has now repeated within a month a Chinese model that is nipping at the heels of American frontier labs more than ever before. Another aggressive price point, another open weights pledge. If Alibaba delivered those weights this week like it was expected to, and you can access them now, the largest open model in history is sitting on Hugging Face for anyone to download, inspect, fine tune and deploy on your own terms. Whatever your view on the geopolitics, for those of us building AI applications, the cost of experimenting at the frontier keeps falling and the control we've retained over our own stacks keeps rising. That doesn't sound like a bad thing to me.

10:10 And finally, as I sometimes do on Friday episodes or at the end of Friday episodes, it's time for a recent Apple Podcast review. Somebody named Civil Service IT said that the show is their go-to favorite. They gave it a five-star rating and said that they've been listening to this podcast for a year or so. They're not an engineer, but they learn so much! That's cool. I'm glad to hear it. I actually, based on some LinkedIn correspondence, I think I know who you are, listener. So thanks to you and thanks to everyone for all the recent ratings and feedback on Apple Podcasts, Spotify, and all the other podcasting platforms out there, as well as for your likes and comments on our YouTube videos. If you can do that for us, if you can provide ratings wherever you listen to your podcasts, please do that. It is the most helpful thing that you can do for me as a listener.

11:02 And bonus points, if you leave written feedback, if you do that, I'll be sure to read your feedback on air like I did today. That seems to be something that's mostly limited to Apple Podcasts. And I think I only see feedback from people that write in the US, but at some point it's on my to-do list to go check other key countries amongst our listenership and read some reviews from there as well. All right, that's the end of today's episode. If you enjoyed it or know someone who might consider sharing this episode with them, tag me in a LinkedIn post with your thoughts. And if you aren't already, be sure to subscribe to the show. Most importantly, however, I just hope you'll keep on listening. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.