Qwen's Architect Just Walked Out the Door

March 5, 2026 · by Mark Bartlett

Download this post as PDF

📚 **Related:** Best Local Models for OpenClaw · Best Local Coding Models 2026 · Qwen 3.5 9B Setup Guide

On March 3rd, Junyang Lin posted six words on X: “me stepping down. bye my beloved qwen.”

Fourteen minutes later, team member Chen Cheng posted: “I know leaving wasn’t your choice.”

Lin was the technical lead and public face of Qwen, Alibaba’s open-weight model family. He joined Alibaba in 2019 and became part of the Qwen team in April 2023. In the time since, he steered Qwen from a lab experiment into the most downloaded open model family on HuggingFace. Over 700 million downloads. Nearly 400 models released. More than 180,000 community fine-tunes built on top.

He’s not the only one out. Yu Bowen, who led post-training, left the same day. Hui Binyuan, who led Qwen Code, had already left for Meta back in January. Three senior departures in three months. The head, the post-training lead, and the coding lead, all gone.

Why It Happened

The reporting from TechCrunch, Bloomberg, and others paints a consistent picture. This wasn’t about money or a better offer. It was a structural disagreement about how to build AI.

Under Lin, the Qwen team operated as a vertically integrated unit. Pre-training, post-training, infrastructure, multimodal, code — all one team, working end-to-end. Lin argued repeatedly that these functions should be tightly coupled. His view was that the people building the training pipeline need to talk to the people doing post-training, who need to talk to the people doing code, who need to see the full picture. One team, one vision.

Alibaba’s Tongyi Lab had a different idea. They wanted to split Qwen into horizontal slices: separate teams for pre-training, post-training, text, multimodal, code. Each reporting up differently. Lin’s management scope would shrink. The vertically integrated structure he’d built would be taken apart.

According to Geopolitechs’ analysis, the commercial pressure is real. Alibaba committed RMB 380 billion to AI infrastructure and swung to a $2.6 billion quarterly free-cash-flow outflow.

Capital at that scale demands returns measured in quarters, not years. There's internal tension about whether open-sourcing models undermines API revenue.

Lin chose to leave rather than watch his team get broken up. CEO Eddie Wu approved the resignation on March 5th and said "I should have known about this earlier." Zhou Jingren, Alibaba Cloud's CTO, takes over. They've hired Zhou Hao from Google DeepMind to lead post-training.

Why This Matters for Local AI

If you're running a local LLM right now, there's a good chance it's a Qwen model.

Qwen 3.5 9B is our default recommendation for 8GB cards. Qwen 2.5 Coder 32B is the best local coding model at 24GB. Qwen 3 32B is the proven OpenClaw agent workhorse. The 27B dense model matches GPT-5 mini on SWE-bench. The small models punch absurdly above their weight.

No other open model family covers the range that Qwen covers. Meta's Llama is strong at the top end but has gaps at the bottom. Mistral releases fewer models and leans commercial. DeepSeek produces excellent reasoning models but doesn't maintain a complete lineup the way Qwen does, from 0.8B to 397B with code variants and MoE variants at every stop.

The Qwen team didn't just make big models and call it a day. They made a 0.8B, a 2B, a 4B, a 9B, a 27B. Code variants. MoE variants. A 35B-A3B that runs agent work at 3B inference speed. They designed the small models with Gated Deltanet so the KV cache stays small at long context. They thought about what runs on real hardware, in real people's homes, and they optimized for it.

That care came from somewhere. It came from a team with a specific vision, led by a specific person.

What Alibaba Says

The official line: Qwen is not being downsized. This is an expansion. They're investing more, hiring more, adding resources.

CEO Eddie Wu apologized for poor communication and said Qwen was "always his first priority." Alibaba said they would "further scale up investment in AI research and development" while "continuing to uphold our open-source model strategy."

I want to believe this. And some of the moves are encouraging. Hiring Zhou Hao from DeepMind is a serious get, over 14,000 Google Scholar citations, contributed to Gemini 3.0. That's not a token hire. Having the CTO directly oversee the lab says they're paying attention at the top.

But.

The team that built Qwen 3.5 is not the team that will build what comes next. The vertically integrated structure that produced 700 million downloads worth of models has been dismantled into horizontal slices. And there's an open question about whether Alibaba's commercial pressure will eventually squeeze the open-source strategy. When you're burning $2.6 billion a quarter, "open source as a community good" starts losing arguments to "how do we monetize this?"

What Realistically Comes Next

The models that exist today aren't going anywhere. Qwen 3.5 is released and on HuggingFace. Your Ollama pulls still work. The 180,000 community fine-tunes don't evaporate.

For the near term, the next 6-12 months, inertia is a powerful thing. There's a pipeline already in motion. Qwen 4, or whatever comes next, was likely already in development before Lin left. The new leadership inherits working infrastructure and training recipes. People who know how things work are still there. It won't fall apart overnight.

The medium term is where I worry. The specific thing that made Qwen special wasn't just compute or data. It was the judgment calls. Which model sizes to release. How much to optimize for small hardware. Whether to publish the 9B or skip straight to the 27B. Whether to use Gated Deltanet in the small models or save it for the flagships. Those decisions came from people who cared about local deployment, and those people are gone.

Alibaba has the resources to keep shipping models. They may ship very good ones. But "very good large model from a Chinese tech giant" is a different thing than "the open model family that actually thinks about what fits on your 8GB card."

The Broader Picture

This is the risk with open source that depends on a single corporate sponsor. Qwen was never a community project. It was an Alibaba project that happened to be open-weight. The community

downloaded it, fine-tuned it, built tools around it. But the decisions about what to build and how were always made inside Alibaba.

When the corporate priorities shift, the community has no seat at that table. We're downstream. We take what they give us and build on it.

This is also why it matters that other open model families exist. Llama 4 Scout is real and running. DeepSeek keeps pushing. Mistral ships. Google open-sourced Gemma. None of them individually fill the Qwen-shaped hole, but collectively they mean that no single departure can kill local AI.

If you're on Qwen today, there's no reason to switch. The models work. They'll keep working. But it's worth paying attention to what Alibaba ships next and whether the small-model focus survives the reorganization. If the next Qwen release is a 70B flagship and a 27B and nothing below that, we'll know what changed.

For Now

Junyang Lin built something that mattered. A model family that runs on hardware regular people can afford. That takes technical skill, but it also takes a kind of stubbornness about who you're building for. Not every AI lab has that. Most don't.

"bye my beloved qwen" is a strange thing to read on a Tuesday morning. I hope wherever he goes next, he keeps building for the rest of us.
