Qwen3.8-Flash-Next: Alibaba's 180B Open Model Beats Bigger Rivals at Coding With Only 6B Active

Kabui, Charles

2026-09-08

Read at ToKnow.ai

On August 26 Alibaba's Qwen team opened the weights of Qwen3.8-Flash-Next, an experimental model that previews the architecture planned for Qwen4. It is a mixture-of-experts model of 180B total parameters, a 125B main model plus 51B of n-gram lookup parameters

and a small prediction head, yet only 6B are active for each token, and it can also read images and video. On Alibaba's own agent-harness runs it scores 62.5 on SWE-bench Pro, ahead of Claude Opus 4.6 Max (53.4), DeepSeek V4 Flash (56.0) and Qwen's larger Qwen3.7-Plus (55.8).

Most capacity lives in n-gram embeddings that can sit in ordinary memory instead of on the accelerator, and attention mixes three linear blocks that compress what came before with one sparse block that reads the most relevant parts of the context. Context is 262,144 tokens natively and can extend toward a million. Alibaba says training took about a ninth of Qwen3.7-Plus's cost, while the hosted version of the same design, Qwen3.8-Flash, costs $0.15 per million input tokens and $0.47 per million output tokens.

The scores are Alibaba's own and the weights carry Qwen's Community License 1.0, free for most uses but requiring a separate deal to sell model-as-a-service access. Still, the direction is clear: efficiency is coming from adding cheap, offloadable capacity and skipping redundant computation rather than scaling one dense transformer, a post-attention turn visible across open-weight releases.

Read More: the dense sibling model that runs on one GPU, Qwen3.8-27B.

Sources:

- Qwen3.8-Flash-Next: A New Architecture, Towards Ultimate Cost-Efficiency (Qwen blog)
- Qwen3.8-Flash-Next model card (Hugging Face)
- Qwen3.8-Flash-Next technical report (GitHub)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*