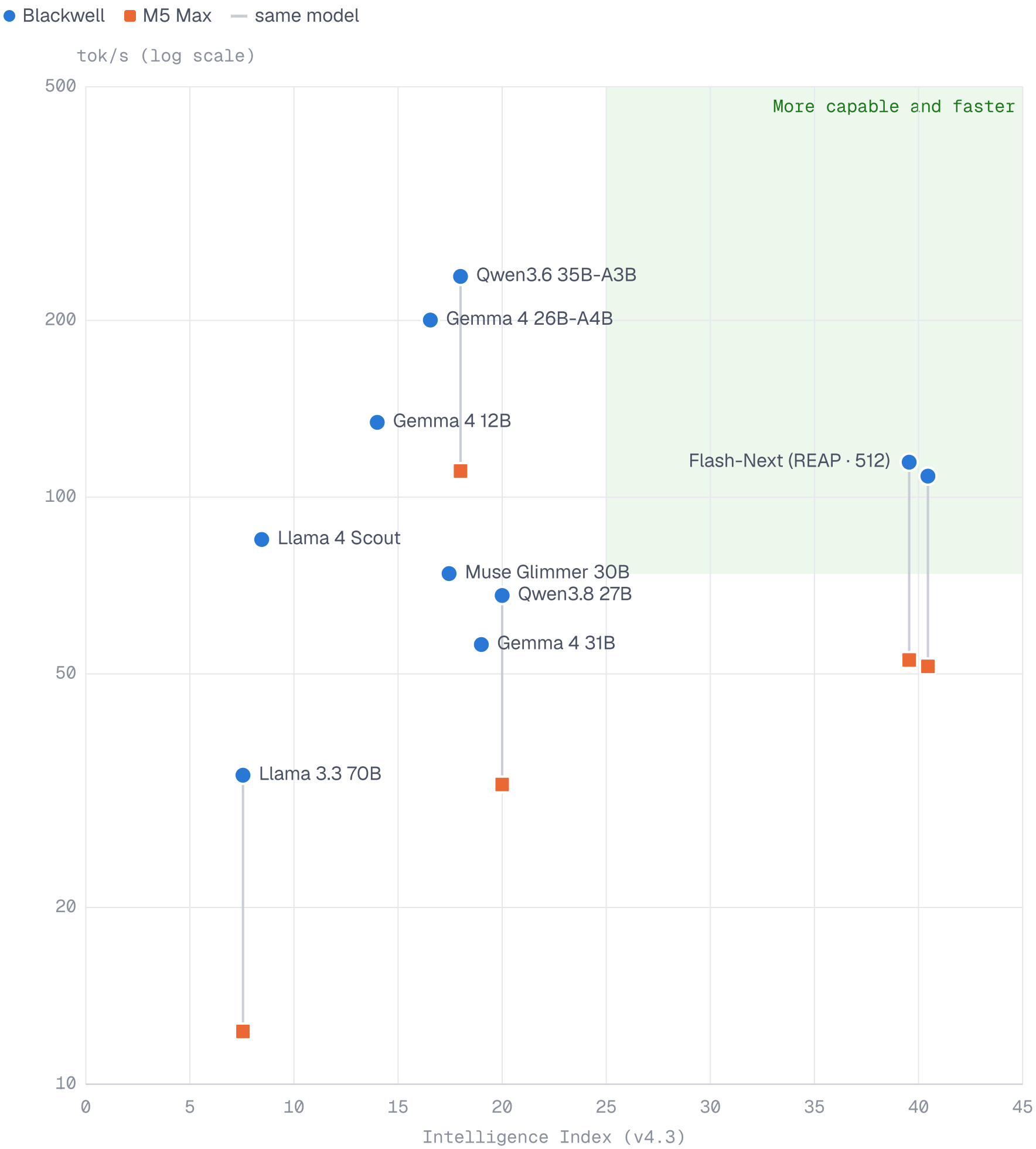
Speed vs intelligence

Top right is better

Output tokens/s for one user (B=1) vs Artificial Analysis Intelligence Index

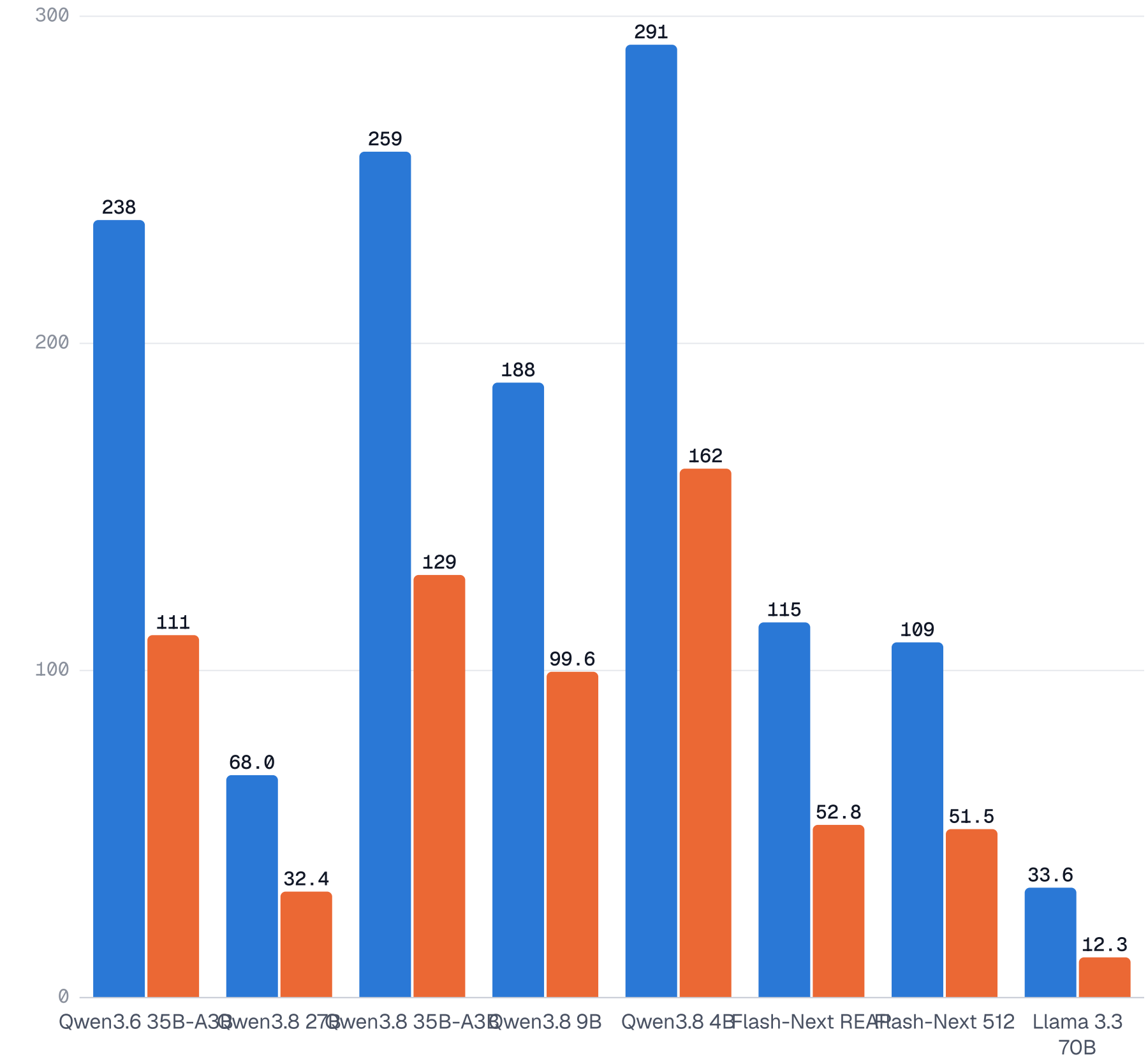


Intelligence: Artificial Analysis Intelligence Index v4.3, read 25 Sep 2026. The score belongs to the base model and does not measure the quantised build we ran. Qwen3.8-27B uses its default entry (26–34 with reasoning effort). Flash-Next ran as a pruned 4-bit build (REAP-288) and as a 3-bit build (512); both are plotted at the base-model score. Six benchmarked models have no index score and are left out: the Qwen3.8 4B, 9B and 35B-A3B distills, Qwen3.6-27B, Laguna XS 2.1 and DeepSeek V4-Flash.

Output speed, one user

Higher is better

Output tokens/s at B=1, models run on both machines, current Qwen models first

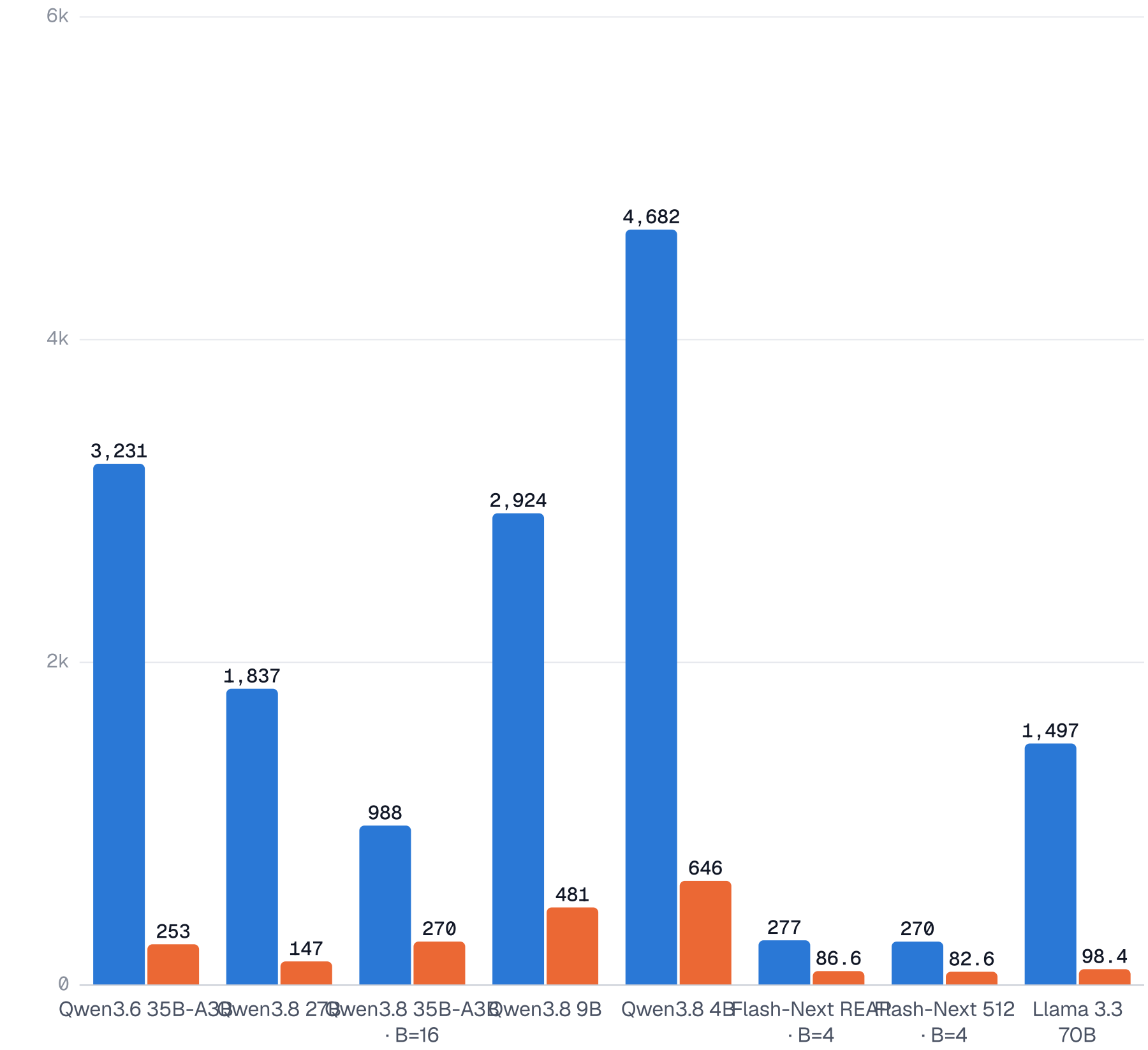


M5 Max ÷ Blackwell ranges from 0.37× to 0.56×. Each bar uses the faster engine on that machine (hover to see which). Qwen3.6 35B-A3B ran on the Mac only through MLX.

Throughput with many users

Higher is better

Total output tokens/s across all concurrent users, at the highest B both machines ran

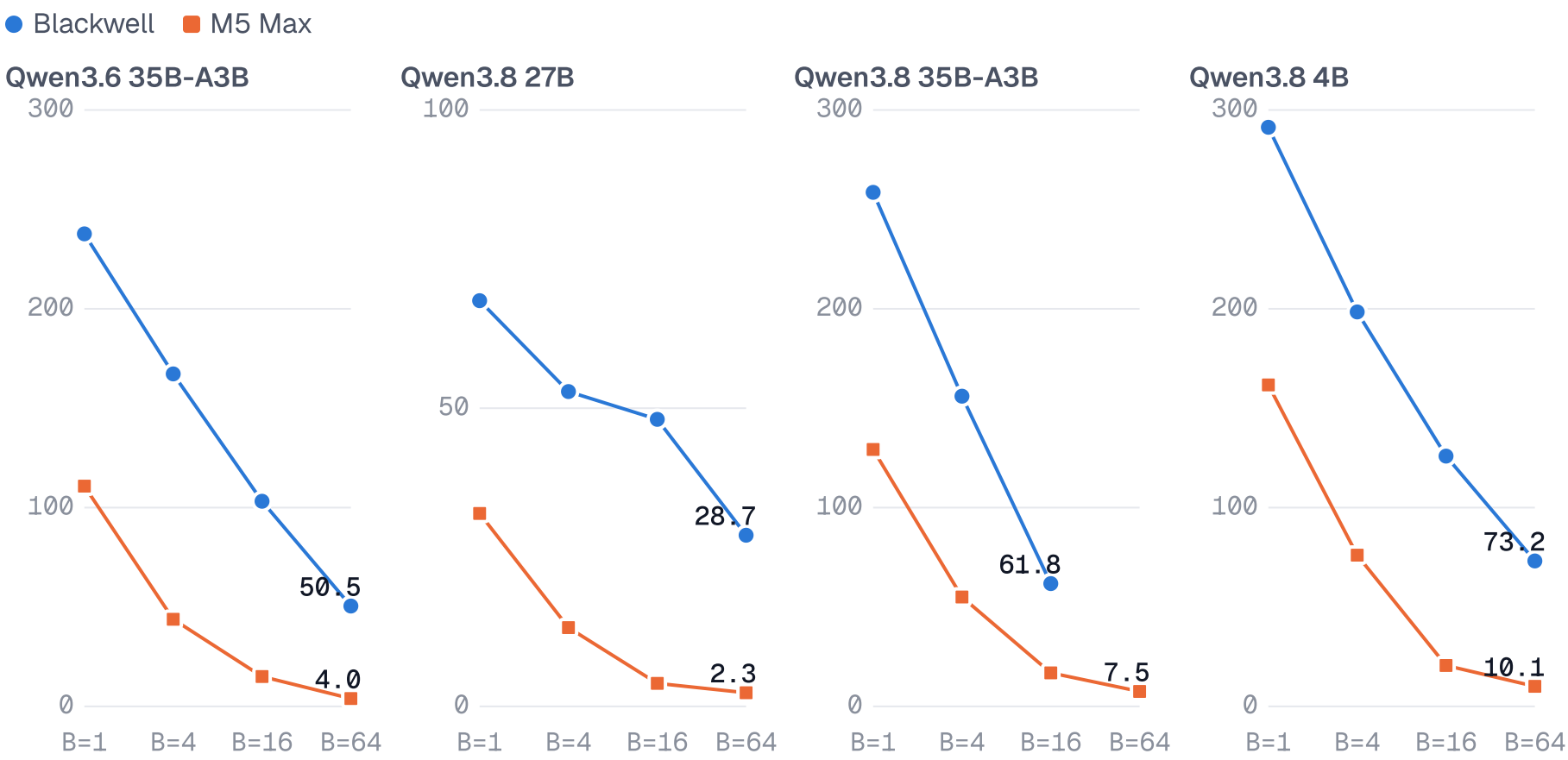


Blackwell's lead comes mostly from vLLM with NVFP4 weights and continuous batching. With the llama.cpp setting, both machines run the same engine. M5 Max MLX figures are conservative bounds.

What each user gets as load grows

Higher is better

Output tokens/s per user, B = 1 → 64



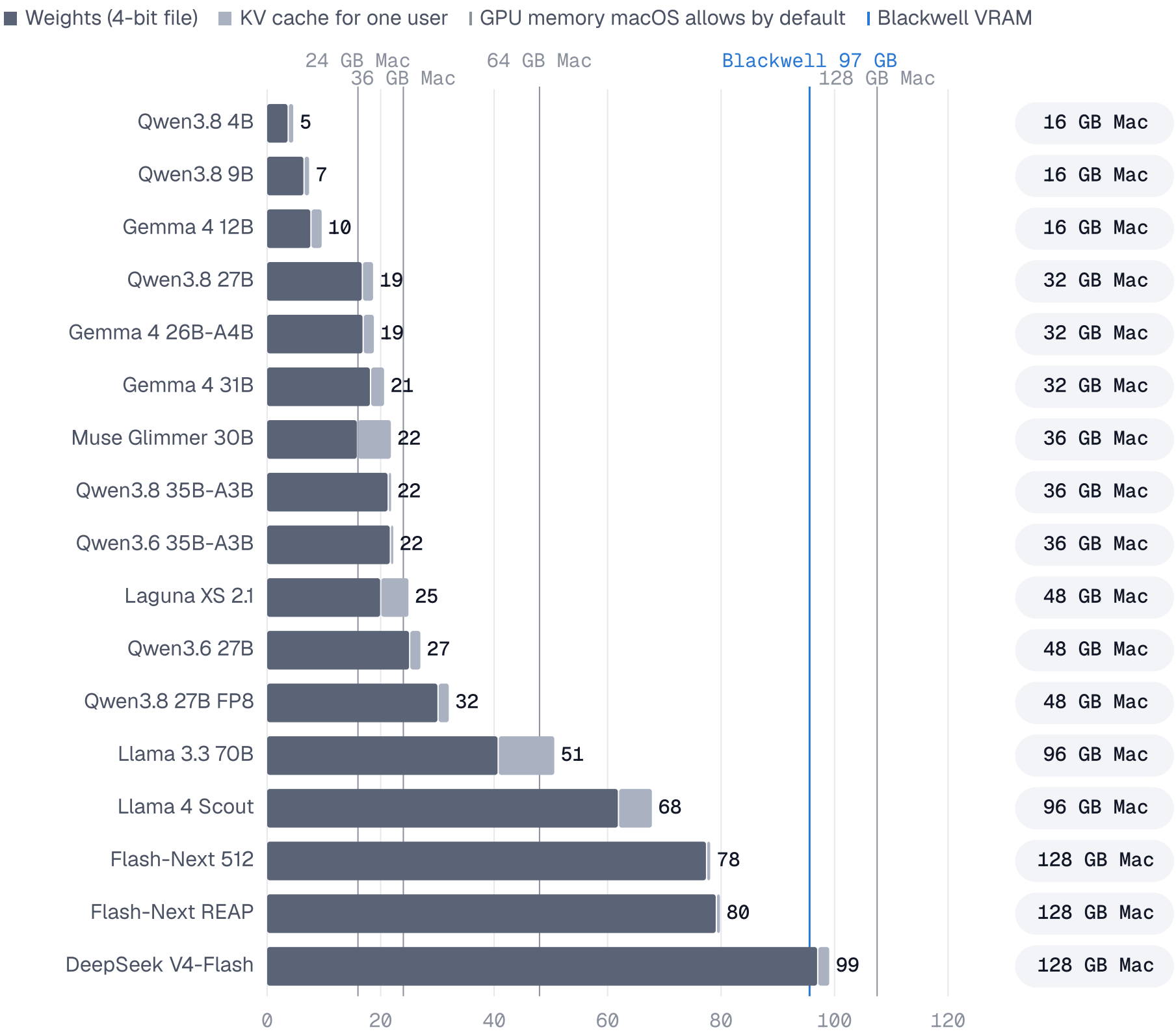
Reading level: about 5–10 tok/s. At B=64, every model on the M5 Max falls to that level or below.

1024-token prompt, 256 tokens out · llama.cpp, vLLM, MLX · Intelligence: Artificial Analysis Index v4.3

Memory needed, and the Mac that fits it

Smaller is better

GiB needed at 32k tokens of context for one user, with the smallest Apple Silicon Mac whose default GPU memory limit fits it



Footprint = weights + KV cache at the selected context + about 1 GiB of runtime buffers. By default macOS lets the GPU use about ⅔ of RAM on Macs up to 36 GB and ¾ above that; the 128 GB M5 Max we tested allowed 107.5 GiB. You can raise the limit with `sudo sysctl iogpu.wired_limit_mb`. KV sizes are estimates from each model family's architecture; only Llama's come from published configs. The Mac RAM verdicts barely depend on them, except for the 70B dense model at long context.