ModelIndex Report Q3 2026

Local LLM performance on consumer hardware -- public summary

EXECUTIVE SUMMARY

Q3 2026: 31 models ranked across 1 hardware tier, measured by 0 active installations on 1 hardware configuration.

Highest score: Qwen3-Coder-30B-A3B-Instruct (IQ1_M) at 69.1 on Apple M4 Max · 128GB+ · 40-core GPU.

On the most-measured tier, Apple M4 Max · 128GB+ · 40-core GPU, Qwen3-Coder-30B-A3B-Instruct (IQ1_M) leads.

Quantization sweet spots: FP16 on 128GB+.

This is the first quarterly report; trends start with the next one.

32 of 34 reports received this quarter carried a verified software signature (94.1%).

TOP MODELS PER HARDWARE TIER

Apple M4 Max · 128GB+ · 40-core GPU

#	Model	Quant	Score	tok/s	Installs
1	Qwen3-Coder-30B-A3B-Instruct	IQ1_M	69.1	81.9	0
2	Gemma 4 12B It	FP16	50.0	--	0
3	NVIDIA-Nemotron-3.5-Lightning-30B-A3B	Q8_K_XL	49.3	16.4	0
4	Qwen-AgentWorld-35B-A3B	FP16	47.4	15.8	0
5	NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16.i1	Q4_0	42.8	9.0	0
6	nemotron-3.5-lightning-30b-a3b	BF16	40.6	8.8	0
7	Qwen3-Coder-30B-A3B-Instruct	Q8_K_XL	39.0	8.3	0
8	NVIDIA-Nemotron-3.5-Lightning-30B-A3B	IQ4_NL	38.2	21.7	0
9	NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16.i1	Q6_K	38.0	35.9	0
10	Qwen3-Coder-Next	Q4_K_M	35.2	12.9	0

QUANTIZATION SWEET SPOTS PER RAM TIER

RAM tier	Quantization	Avg score	Avg tok/s	Rows
128GB+	FP16	48.7	15.8	2

TOP MODEL PAIRS PER TIER

TRENDS

This is the first quarterly report; trends start with the next one.

DATA PROVENANCE

Active installations: 0. Hardware configurations: 1. Reports received this quarter: 34, of which 32 carried a verified software signature (94.1%).

METHODOLOGY

Every figure is computed from anonymous performance reports that RunMyAI Desktop installations send automatically (Terms 6.9). A report carries model, quantization and hardware-specification measurements only -- no content, no identity.

Only reports signed with the RunMyAI software-attestation key are ranked; the signature proves a genuine build produced the bytes.

Reports are grouped by hardware tier: processor model, RAM tier and accelerator. A model is ranked on a tier from the measurements of every installation that ran it there.

Speed outliers are rejected per model, quantization and tier with a modified z-score around the median (cut-off 3.5) before averaging.

The 0-100 score weights delivered accuracy, sustained speed, time to first token and user ratings, renormalized over what was measured and capped by how much of it was measured.

Quantization sweet spot: the quantization with the highest average score among the ranked rows of a RAM tier.

Trends compare each ranked row with the same model, quantization and tier in the previous quarterly report.