GPUpartner

VALIDATION REPORT · TEXT GENERATION

Phi 3.5 MOE Instruct

on M4 Max 128 GB · 8-bit (Q8_0) · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

Not validated

4 of 5 checks pass

Phi 3.5 MOE Instruct at 8-bit runs on the M4 Max 128 GB with 128K of context at about 46 tokens a second.

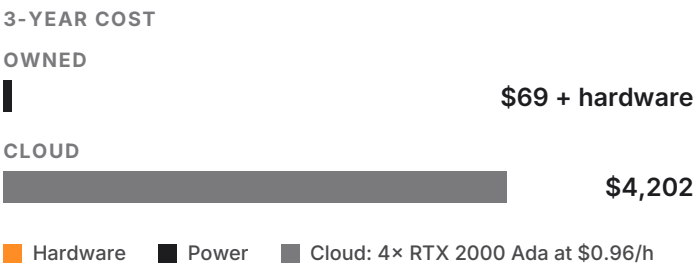
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	96.0 GB usable	57.9 GB needed	PASS
Headroom	5% free	38.1 GB free (40%)	PASS
Context for chat	16K or more	128K	PASS
Writing speed	15 tok/s or more	46 tok/s	PASS
First reply	under 3 s, 2K prompt	7.6 s	FAIL

Targets: a reply you can read along with, a first reply the use case tolerates, and 5% of memory left free. Calculated from published specs, not benchmarked.

BUDGET

HARDWARE, ONE-TIME	MSRP not entered
POWER, PER YEAR	$23 · 87 W at the wall
CLOUD INSTEAD, PER YEAR	$1,401
3-YEAR TOTAL, OWNED	Hardware + $69 power
3-YEAR TOTAL, CLOUD	$4,202
BREAK-EVEN	Needs the hardware MSRP



Budget basis: 4 h a day, power at $0.18 per kWh, cloud at RunPod on-demand prices as of Nov 22, 2026. Hardware at the MSRP entered in the calculator, not a GPU Partner quote. A 3-year view; resale value, staff time and taxes aren't counted.

GPUpartner

TECHNICAL DETAIL · TEXT GENERATION

Phi 3.5 MOE Instruct

on M4 Max 128 GB · 8-bit (Q8_0) · Ollama / LM Studio · Chat

SETUP

PARAMETERS	41.9B, 6.6B active
PRECISION	8-bit (Q8_0)
CONTEXT	128K, auto
CHATS AT ONCE	1
USE	Chat
RUNS IN	Ollama / LM Studio
HARDWARE	M4 Max 128 GB
GPU CAN USE	96.0 GB of 128 GB

MEMORY

WEIGHTS	41.4 GB
CONTEXT	16.0 GB
RUNTIME	447 MB
SYSTEM	32.0 GB
FREE	38.1 GB

SAME MODEL, OTHER MACHINES

RTX 4090	3-bit · 153 tok/s
RTX 5090	5-bit · 176 tok/s
RTX PRO 6000 BLACKWELL	8-bit · 133 tok/s
M5 MAX 128 GB	8-bit · 50 tok/s
M3 ULTRA 256 GB	8-bit · 56 tok/s
DGX SPARK 128 GB	8-bit · 26 tok/s
RYZEN AI MAX+ 395 128 GB	8-bit · 20 tok/s

HOW TO RUN IT

LLAMA.CPP	`llama-server -hf bartowski/Phi-3.5-MoE-instruct-GGUF`
VLLM	`vllm serve microsoft/Phi-3.5-MoE-instruct`
MODEL CARD	huggingface.co/microsoft/Phi-3.5-MoE-instruct

METHOD AND ASSUMPTIONS

Memory is the weights (parameters times bits per weight, read from each model's config.json), the context cache for every chat, and the runtime, against the memory this GPU can use after the OS and display take theirs. Writing speed is memory bandwidth over the bytes read per token, plus a per-layer latency, calibrated against llama.cpp benchmarks. First reply is the prompt's FLOPs over the chip's dense 16-bit TFLOPS at the runtime's measured efficiency. Running cost is power at the wall at your rate and hours; the cloud figure is the cheapest RunPod on-demand GPU that runs the same job. Accuracy: memory lands within about a gigabyte; speeds and times are rough, about ±25%. Model data and hardware specs checked Nov 2026; cloud prices are RunPod's on-demand list as of Nov 22, 2026. Report VR-05HSLT1, generated November 22, 2026.