GPUpartner

VALIDATION REPORT · TEXT GENERATION

Llama 3.1 70B

on M4 Pro 64 GB · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

Not validated

3 of 5 checks pass

Llama 3.1 70B at 4-bit runs on the M4 Pro 64 GB with 16K of context at about 5.1 tokens a second.

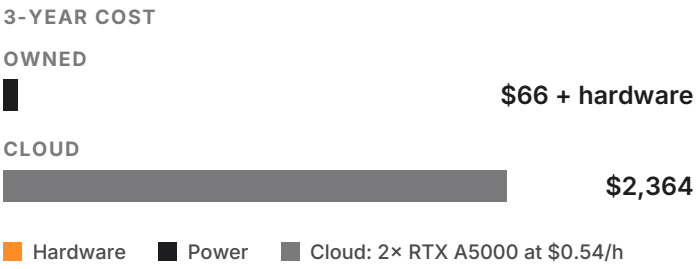
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	48.0 GB usable	45.6 GB needed	PASS
Headroom	5% free	2.4 GB free (5%)	PASS
Context for chat	16K or more	16K	PASS
Writing speed	15 tok/s or more	5.1 tok/s	FAIL
First reply	under 3 s, 2K prompt	48 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	$22 · 84 W at the wall
CLOUD INSTEAD, PER YEAR	$788
3-YEAR TOTAL, OWNED	Hardware + $66 power
3-YEAR TOTAL, CLOUD	$2,364
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

Llama 3.1 70B

on M4 Pro 64 GB · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

SETUP

PARAMETERS	70.6B
PRECISION	4-bit (Q4_K_M)
CONTEXT	16K, auto
CHATS AT ONCE	1
USE	Chat
RUNS IN	Ollama / LM Studio
HARDWARE	M4 Pro 64 GB
GPU CAN USE	48.0 GB of 64.0 GB

MEMORY

WEIGHTS	39.8 GB
CONTEXT	5.0 GB
RUNTIME	763 MB
SYSTEM	16.0 GB
FREE	2.4 GB

SUGGESTED CHANGES

8-BIT CACHE	Room for 28K context

SAME MODEL, OTHER MACHINES

RTX 4090	Too big
RTX 5090	Too big
RTX PRO 6000 BLACKWELL	8-bit · 16 tok/s
M5 MAX 128 GB	8-bit · 6.0 tok/s
M3 ULTRA 256 GB	8-bit · 6.9 tok/s
DGX SPARK 128 GB	8-bit · 2.6 tok/s
RYZEN AI MAX+ 395 128 GB	8-bit · 2.1 tok/s

HOW TO RUN IT

LLAMA.CPP	`llama-server -hf NousResearch/Hermes-3-Llama-3.1-70B-GGUF`
VLLM	`vllm serve meta-llama/Llama-3.1-70B`
MODEL CARD	huggingface.co/meta-llama/Llama-3.1-70B

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-1QJK8QS, generated November 22, 2026.