GPUpartner

VALIDATION REPORT · TEXT GENERATION

DeepSeek V4 Pro

on M4 Max 128 GB · FP4 + FP8 · Ollama / LM Studio · Chat

STATUS · OCT 22, 2026

Not validated

0 of 2 checks pass

DeepSeek V4 Pro at FP4 + FP8 needs 806 GB; the M4 Max 128 GB can use 96.0 GB. The smallest machine we checked that runs it is the H200 NVL.

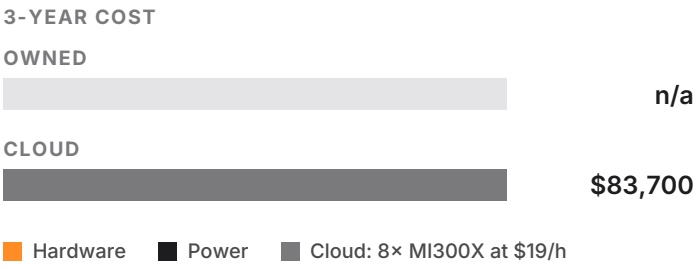
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	96.0 GB usable	806 GB needed	FAIL
Headroom	5% free	none	FAIL
Context for chat	16K or more	–	NOT TESTED
Writing speed	15 tok/s or more	–	NOT TESTED
First reply	under 3 s, 2K prompt	–	NOT TESTED

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	Doesn't run on this machine
CLOUD INSTEAD, PER YEAR	$27,900
3-YEAR TOTAL, OWNED	n/a
3-YEAR TOTAL, CLOUD	$83,700
BREAK-EVEN	n/a



Budget basis: 4 h a day, power at $0.18 per kWh, cloud at RunPod on-demand prices as of Oct 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

DeepSeek V4 Pro

on M4 Max 128 GB · FP4 + FP8 · Ollama / LM Studio · Chat

SETUP

PARAMETERS	1.6T, 49B active
PRECISION	FP4 + FP8
CONTEXT	8K, 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	805 GB
CONTEXT	77 MB
RUNTIME	733 MB
SYSTEM	32.0 GB
OVER BY	710 GB

SUGGESTED CHANGES

8× H200 NVL	Runs it with 1M context

SAME MODEL, OTHER MACHINES

RTX 4090	Too big
RTX 5090	Too big
RTX PRO 6000 BLACKWELL	Too big
M5 MAX 128 GB	Too big
M3 ULTRA 256 GB	Too big
DGX SPARK 128 GB	Too big
RYZEN AI MAX+ 395 128 GB	Too big

HOW TO RUN IT

LLAMA.CPP	`llama-server -hf unsloth/DeepSeek-V4-Pro-0813-GGUF`
VLLM	`vllm serve deepseek-ai/DeepSeek-V4-Pro`
MLX (MAC)	`mlx_lm.chat --model mlx-community/DeepSeek-V4-Pro-4bit`
MODEL CARD	huggingface.co/deepseek-ai/DeepSeek-V4-Pro

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 Oct 2026; cloud prices are RunPod's on-demand list as of Oct 22, 2026. Report VR-12WB0V1, generated October 22, 2026.