GPUpartner

VALIDATION REPORT · TEXT GENERATION

Kimi K3

on 4× H200 NVL · INT4 · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

Not validated

0 of 2 checks pass

Kimi K3 at INT4 needs 1,457 GB; the 4× H200 NVL can use 564 GB.

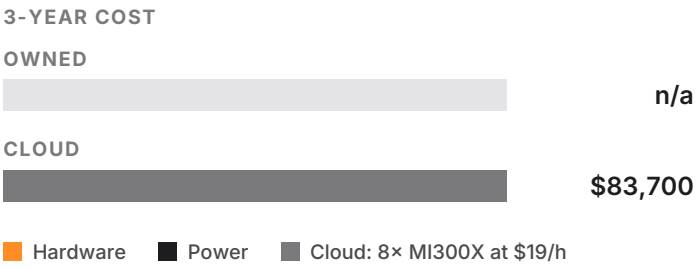
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	564 GB usable	1,457 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 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

Kimi K3

on 4× H200 NVL · INT4 · Ollama / LM Studio · Chat

SETUP

PARAMETERS	2.8T, 104B active
PRECISION	INT4
CONTEXT	8K, auto
CHATS AT ONCE	1
USE	Chat
RUNS IN	Ollama / LM Studio
HARDWARE	4× H200 NVL
GPU CAN USE	564 GB of 564 GB

MEMORY

WEIGHTS	1,454 GB
CONTEXT	645 MB
RUNTIME	3.1 GB
SYSTEM	0.0 GB
OVER BY	893 GB

SUGGESTED CHANGES

1.6-BIT WEIGHTS	Fits on the GPU, 8K 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/Kimi-K3-GGUF`
VLLM	`vllm serve moonshotai/Kimi-K3`
MODEL CARD	huggingface.co/moonshotai/Kimi-K3

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-0TQSI5V, generated November 22, 2026.