GPUpartner

VALIDATION REPORT · TEXT GENERATION

GLM 4.1V 9B Thinking

on RTX 3070 · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

Validated

5 of 5 checks pass

GLM 4.1V 9B Thinking at 4-bit runs on the RTX 3070 with 20K of context at about 50 tokens a second.

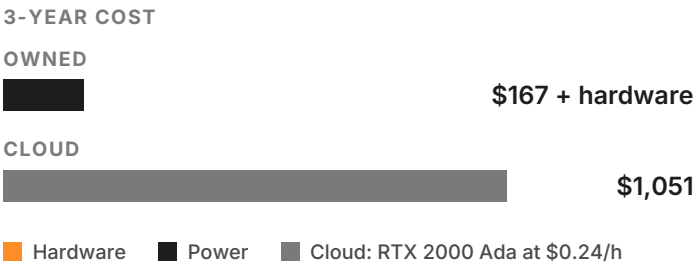
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	8.0 GB usable	7.5 GB needed	PASS
Headroom	5% free	556 MB free (7%)	PASS
Context for chat	16K or more	20K	PASS
Writing speed	15 tok/s or more	50 tok/s	PASS
First reply	under 3 s, 2K prompt	1.0 s	PASS

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	$56 · 212 W at the wall
CLOUD INSTEAD, PER YEAR	$350
3-YEAR TOTAL, OWNED	Hardware + $167 power
3-YEAR TOTAL, CLOUD	$1,051
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

GLM 4.1V 9B Thinking

on RTX 3070 · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

SETUP

PARAMETERS	10.3B
PRECISION	4-bit (Q4_K_M)
CONTEXT	20K, auto
CHATS AT ONCE	1
USE	Chat
RUNS IN	Ollama / LM Studio
HARDWARE	RTX 3070
GPU CAN USE	8.0 GB of 8.0 GB

MEMORY

WEIGHTS	5.8 GB
CONTEXT	800 MB
RUNTIME	885 MB
SYSTEM	0.0 GB
FREE	556 MB

SUGGESTED CHANGES

8-BIT CACHE	Room for 44K context

SAME MODEL, OTHER MACHINES

RTX 4090	8-bit · 63 tok/s
RTX 5090	8-bit · 106 tok/s
RTX PRO 6000 BLACKWELL	8-bit · 106 tok/s
M5 MAX 128 GB	8-bit · 39 tok/s
M3 ULTRA 256 GB	8-bit · 44 tok/s
DGX SPARK 128 GB	8-bit · 18 tok/s
RYZEN AI MAX+ 395 128 GB	8-bit · 15 tok/s

HOW TO RUN IT

LLAMA.CPP	`llama-server -hf unsloth/GLM-4.1V-9B-Thinking-GGUF`
VLLM	`vllm serve zai-org/GLM-4.1V-9B-Thinking`
MLX (MAC)	`mlx_lm.chat --model mlx-community/GLM-4.1V-9B-Thinking-8bit`
MODEL CARD	huggingface.co/zai-org/GLM-4.1V-9B-Thinking

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-0CB5XCL, generated November 22, 2026.