GPUpartner

VALIDATION REPORT · TEXT GENERATION

GLM-5.3 Flash

on DGX Spark 128 GB · 2-bit (Q2_K) · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

Validated with conditions

4 of 5 checks pass

GLM-5.3 Flash at 2-bit runs on the DGX Spark 128 GB with 104K of context at about 26 tokens a second.

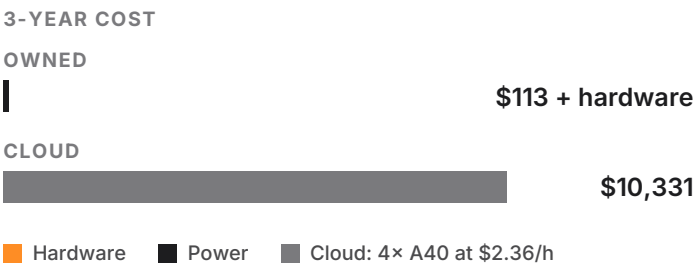
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	120 GB usable	114 GB needed	PASS
Headroom	5% free	6.0 GB free (5%)	PASS
Context for chat	16K or more	104K	PASS
Writing speed	15 tok/s or more	26 tok/s	PASS
First reply	under 3 s, 2K prompt	3.9 s	CAUTION

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	$38 · 144 W at the wall
CLOUD INSTEAD, PER YEAR	$3,444
3-YEAR TOTAL, OWNED	Hardware + $113 power
3-YEAR TOTAL, CLOUD	$10,331
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-5.3 Flash

on DGX Spark 128 GB · 2-bit (Q2_K) · Ollama / LM Studio · Chat

SETUP

PARAMETERS	321B, 18B active
PRECISION	2-bit (Q2_K)
CONTEXT	104K, auto
CHATS AT ONCE	1
USE	Chat
RUNS IN	Ollama / LM Studio
HARDWARE	DGX Spark 128 GB
GPU CAN USE	120 GB of 128 GB

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	3-bit · 43 tok/s
RYZEN AI MAX+ 395 128 GB	Too big
DGX STATION GB300	5-bit · 188 tok/s

MEMORY

WEIGHTS	112 GB
CONTEXT	1.4 GB
RUNTIME	687 MB
SYSTEM	7.7 GB
FREE	6.0 GB

HOW TO RUN IT

LLAMA.CPP	`llama-server -hf unsloth/GLM-5.3-Flash-GGUF`
VLLM	`vllm serve zai-org/GLM-5.3-Flash`
MLX (MAC)	`mlx_lm.chat --model mlx-community/GLM-5.3-Flash-4bit`
MODEL CARD	huggingface.co/zai-org/GLM-5.3-Flash

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-1OFQR5Z, generated November 22, 2026.