GPUpartner

VALIDATION REPORT · TEXT GENERATION

Gemma 4 E2B

on DGX Spark 128 GB · 8-bit (Q8_0) · Ollama / LM Studio · Chat

STATUS · OCT 23, 2026

Validated

5 of 5 checks pass

Gemma 4 E2B at 8-bit runs on the DGX Spark 128 GB with 128K of context at about 61 tokens a second.

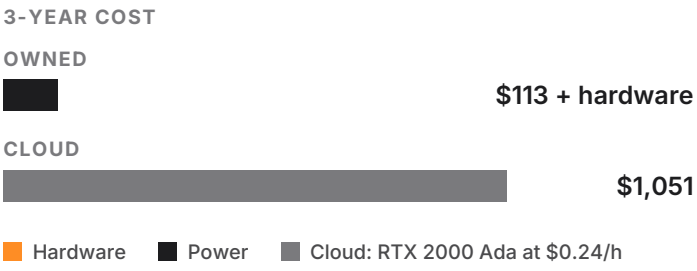
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	120 GB usable	6.6 GB needed	PASS
Headroom	5% free	114 GB free (94%)	PASS
Context for chat	16K or more	128K	PASS
Writing speed	15 tok/s or more	61 tok/s	PASS
First reply	under 3 s, 2K prompt	0.1 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	$38 · 144 W at the wall
CLOUD INSTEAD, PER YEAR	$350
3-YEAR TOTAL, OWNED	Hardware + $113 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 Oct 23, 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

Gemma 4 E2B

on DGX Spark 128 GB · 8-bit (Q8_0) · Ollama / LM Studio · Chat

SETUP

PARAMETERS	5.1B, 2.3B active
PRECISION	8-bit (Q8_0)
CONTEXT	128K, 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	8-bit · 191 tok/s
RTX 5090	8-bit · 294 tok/s
RTX PRO 6000 BLACKWELL	8-bit · 294 tok/s
M5 MAX 128 GB	8-bit · 109 tok/s
M3 ULTRA 256 GB	8-bit · 120 tok/s
RYZEN AI MAX+ 395 128 GB	8-bit · 48 tok/s
DGX STATION GB300	8-bit · 609 tok/s

MEMORY

WEIGHTS	5.1 GB
CONTEXT	774 MB
RUNTIME	816 MB
SYSTEM	7.7 GB
FREE	114 GB

HOW TO RUN IT

OLLAMA	`ollama run gemma4:e2b`
LLAMA.CPP	`llama-server -hf unsloth/gemma-4-E2B-it-GGUF`
VLLM	`vllm serve google/gemma-4-E2B-it`
MLX (MAC)	`mlx_lm.chat --model mlx-community/gemma-4-e2b-it-4bit`
MODEL CARD	huggingface.co/google/gemma-4-E2B-it

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 23, 2026. Report VR-1QN3C8Z, generated October 23, 2026.