GPUpartner

VALIDATION REPORT · TEXT GENERATION

Step 3.5 Flash

on A100 80 GB SXM · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

Not validated

0 of 2 checks pass

Step 3.5 Flash at 4-bit needs 117 GB; the A100 80 GB SXM can use 80.0 GB. The smallest machine we checked that runs it is the H200 NVL.

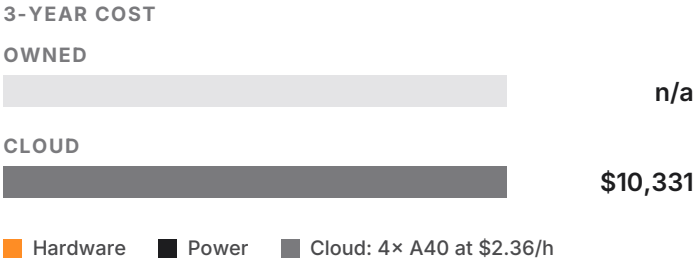
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	80.0 GB usable	117 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	$3,444
3-YEAR TOTAL, OWNED	n/a
3-YEAR TOTAL, CLOUD	$10,331
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

Step 3.5 Flash

on A100 80 GB SXM · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

SETUP

PARAMETERS	199B, 9.4B active
PRECISION	4-bit (Q4_K_M)
CONTEXT	8K, auto
CHATS AT ONCE	1
USE	Chat
RUNS IN	Ollama / LM Studio
HARDWARE	A100 80 GB SXM
GPU CAN USE	80.0 GB of 80.0 GB

MEMORY

WEIGHTS	113 GB
CONTEXT	3.6 GB
RUNTIME	841 MB
SYSTEM	0.0 GB
OVER BY	37.0 GB

SUGGESTED CHANGES

2-BIT WEIGHTS	Fits on the GPU, 13K context
2× A100 80 GB SXM	Fits across the cards, 96K context
H200 NVL	Runs it with 52K context
M5 ULTRA · 256 GB	Runs it with 176K context

SAME MODEL, OTHER MACHINES

RTX 4090	Too big
RTX 5090	Too big
RTX PRO 6000 BLACKWELL	2-bit · 181 tok/s
M5 MAX 128 GB	2-bit · 72 tok/s
M3 ULTRA 256 GB	6-bit · 48 tok/s
DGX SPARK 128 GB	3-bit · 37 tok/s
RYZEN AI MAX+ 395 128 GB	2-bit · 37 tok/s

HOW TO RUN IT

LLAMA.CPP	`llama-server -hf ggml-org/Step-3.5-Flash-GGUF`
VLLM	`vllm serve stepfun-ai/Step-3.5-Flash`
MLX (MAC)	`mlx_lm.chat --model mlx-community/Step-3.5-Flash-4bit`
MODEL CARD	huggingface.co/stepfun-ai/Step-3.5-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-04H5K99, generated November 22, 2026.