GPUpartner

VALIDATION REPORT · TEXT GENERATION

# Nemotron 3 Ultra 550B A55B

on H100 SXM · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

**Not validated**

0 of 2 checks pass

**Nemotron 3 Ultra 550B A55B at 4-bit needs 318 GB; the H100 SXM can use 80.0 GB. The smallest machine we checked that runs it is the RTX PRO 6000 Blackwell.**

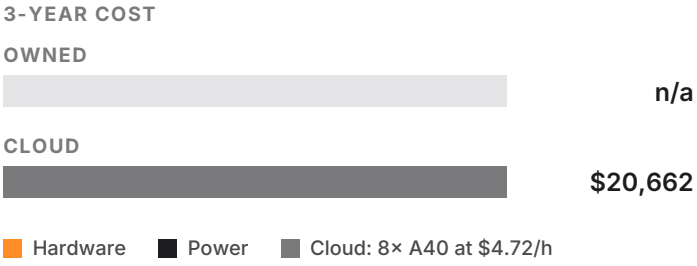
## VALIDATION

| CHECK | TARGET | RESULT | STATUS |
|---|---|---|---|
| Fits in GPU memory | 80.0 GB usable | 318 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 | $6,887 |
| 3-YEAR TOTAL, OWNED | n/a |
| 3-YEAR TOTAL, CLOUD | $20,662 |
| 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

# Nemotron 3 Ultra 550B A55B

on H100 SXM · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

## SETUP

| | |
|---|---|
| PARAMETERS | 561B, 55B active |
| PRECISION | 4-bit (Q4_K_M) |
| CONTEXT | 8K, auto |
| CHATS AT ONCE | 1 |
| USE | Chat |
| RUNS IN | Ollama / LM Studio |
| HARDWARE | H100 SXM |
| GPU CAN USE | 80.0 GB of 80.0 GB |

## MEMORY

| | |
|---|---|
| WEIGHTS | 316 GB |
| CONTEXT | 485 MB |
| RUNTIME | 0.9 GB |
| SYSTEM | 0.0 GB |
| OVER BY | 238 GB |

## SUGGESTED CHANGES

| | |
|---|---|
| 5× H100 SXM | Fits across the cards, 256K context |
| 4× RTX PRO 6000 BLACKWELL | Runs it with 256K context |
| M5 ULTRA · 512 GB | Runs it with 256K context |
| 4× H200 NVL | Runs it with 256K 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 | 1.6-bit · 27 tok/s |
| DGX SPARK 128 GB | 1.6-bit · 15 tok/s |
| RYZEN AI MAX+ 395 128 GB | Too big |

## HOW TO RUN IT

| | |
|---|---|
| LLAMA.CPP | `llama-server -hf unsloth/NVIDIA-Nemotron-3-Ultra-550B-A55B-GGUF` |
| VLLM | `vllm serve nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16` |
| MLX (MAC) | `mlx_lm.chat --model mlx-community/Nemotron-3-Ultra-550B-A55B-4bit` |
| MODEL CARD | huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16 |

## 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-1XHCOTE, generated November 22, 2026.