GPUpartner

VALIDATION REPORT · TEXT GENERATION

# Yi 34B 200K

on RTX 5080 · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

**Not validated**

0 of 2 checks pass

**Yi 34B 200K at 4-bit needs 22.1 GB; the RTX 5080 can use 16.0 GB. The smallest machine we checked that runs it is the RTX 5090.**

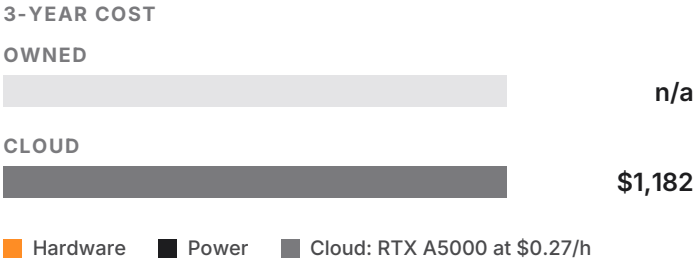
## VALIDATION

| CHECK | TARGET | RESULT | STATUS |
|---|---|---|---|
| Fits in GPU memory | 16.0 GB usable | 22.1 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 | $394 |
| 3-YEAR TOTAL, OWNED | n/a |
| 3-YEAR TOTAL, CLOUD | $1,182 |
| 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

# Yi 34B 200K

on RTX 5080 · 4-bit (Q4_K_M) · Ollama / LM Studio · Chat

## SETUP

| | |
|---|---|
| PARAMETERS | 34.4B |
| PRECISION | 4-bit (Q4_K_M) |
| CONTEXT | 8K, auto |
| CHATS AT ONCE | 1 |
| USE | Chat |
| RUNS IN | Ollama / LM Studio |
| HARDWARE | RTX 5080 |
| GPU CAN USE | 16.0 GB of 16.0 GB |

## MEMORY

| | |
|---|---|
| WEIGHTS | 19.4 GB |
| CONTEXT | 1.9 GB |
| RUNTIME | 810 MB |
| SYSTEM | 0.0 GB |
| OVER BY | 6.1 GB |

## SUGGESTED CHANGES

| | |
|---|---|
| 2× RTX 5080 | Fits across the cards, 40K context |
| SPILL TO RAM | Runs, slower, 6.0 tok/s |
| RTX 5090 | Runs it with 40K context |
| M5 MAX · 48 GB | Runs it with 60K context |

## SAME MODEL, OTHER MACHINES

| | |
|---|---|
| RTX 4090 | 3-bit · 40 tok/s |
| RTX 5090 | 5-bit · 48 tok/s |
| RTX PRO 6000 BLACKWELL | 8-bit · 33 tok/s |
| M5 MAX 128 GB | 8-bit · 12 tok/s |
| M3 ULTRA 256 GB | 8-bit · 14 tok/s |
| DGX SPARK 128 GB | 8-bit · 5.3 tok/s |
| RYZEN AI MAX+ 395 128 GB | 8-bit · 4.2 tok/s |

## HOW TO RUN IT

| | |
|---|---|
| LLAMA.CPP | `llama-server -hf TheBloke/Yi-34B-200K-GGUF` |
| VLLM | `vllm serve 01-ai/Yi-34B-200K` |
| MODEL CARD | huggingface.co/01-ai/Yi-34B-200K |

## 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-1I9MGOP, generated November 22, 2026.