GPUpartner

VALIDATION REPORT · TEXT GENERATION

# Idefics3 8B Llama3

on RTX 5070 · 8-bit (Q8_0) · Ollama / LM Studio · Chat

STATUS · NOV 22, 2026

**Validated**

5 of 5 checks pass

**Idefics3 8B Llama3 at 8-bit runs on the RTX 5070 with 16K of context at about 53 tokens a second.**

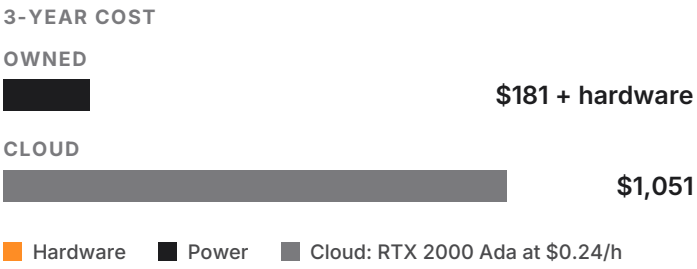
## VALIDATION

| CHECK | TARGET | RESULT | STATUS |
|---|---|---|---|
| Fits in GPU memory | 12.0 GB usable | 11.2 GB needed | PASS |
| Headroom | 5% free | 826 MB free (7%) | PASS |
| Context for chat | 16K or more | 16K | PASS |
| Writing speed | 15 tok/s or more | 53 tok/s | PASS |
| First reply | under 3 s, 2K prompt | 0.5 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 | $60 · 230 W at the wall |
| CLOUD INSTEAD, PER YEAR | $350 |
| 3-YEAR TOTAL, OWNED | Hardware + $181 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 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

# Idefics3 8B Llama3

on RTX 5070 · 8-bit (Q8_0) · Ollama / LM Studio · Chat

## SETUP

| | |
|---|---|
| PARAMETERS | 8.5B |
| PRECISION | 8-bit (Q8_0) |
| CONTEXT | 16K, auto |
| CHATS AT ONCE | 1 |
| USE | Chat |
| RUNS IN | Ollama / LM Studio |
| HARDWARE | RTX 5070 |
| GPU CAN USE | 12.0 GB of 12.0 GB |

## MEMORY

| | |
|---|---|
| WEIGHTS | 8.4 GB |
| CONTEXT | 2.0 GB |
| RUNTIME | 839 MB |
| SYSTEM | 0.0 GB |
| FREE | 826 MB |

## SUGGESTED CHANGES

| | |
|---|---|
| 8-BIT CACHE | Room for 32K context |

## SAME MODEL, OTHER MACHINES

| | |
|---|---|
| RTX 4090 | 8-bit · 77 tok/s |
| RTX 5090 | 8-bit · 130 tok/s |
| RTX PRO 6000 BLACKWELL | 8-bit · 130 tok/s |
| M5 MAX 128 GB | 8-bit · 48 tok/s |
| M3 ULTRA 256 GB | 8-bit · 54 tok/s |
| DGX SPARK 128 GB | 8-bit · 22 tok/s |
| RYZEN AI MAX+ 395 128 GB | 8-bit · 18 tok/s |

## HOW TO RUN IT

| | |
|---|---|
| LLAMA.CPP | `llama-server -hf jesse9527/Idefics3-8B-Llama3-Q4_K_M-GGUF` |
| VLLM | `vllm serve HuggingFaceM4/Idefics3-8B-Llama3` |
| MODEL CARD | huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3 |

## 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-148JIR5, generated November 22, 2026.