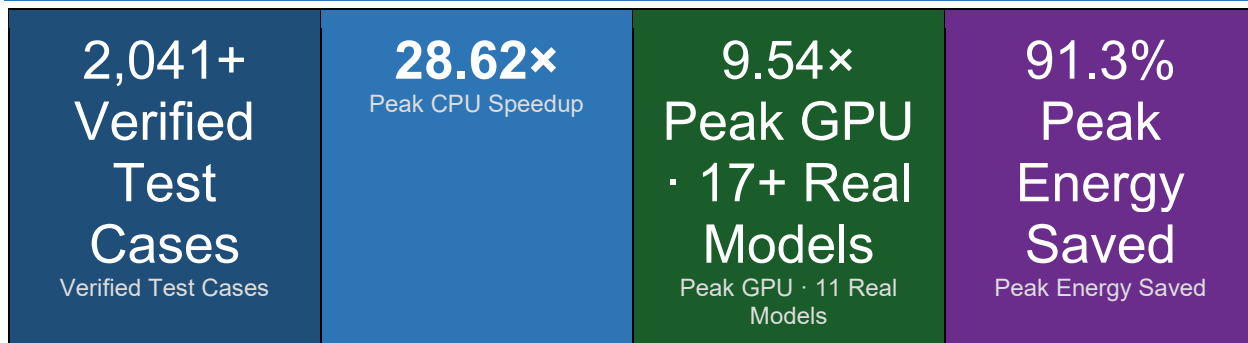


# ROLV Primitive©

## Complete Benchmark Report

Updated: April 7, 2026 · Rolv Eitrem Heggenhougen · rolv.ai · rolv@rolv.ai  
ROLV Primitive© · RSMT™ · ROLVswitch™ · 3 Patents Pending



### 1 · Platform Summary

All results: 4 SHA-256 hashes · perturbation test · ATOL=0.05 · baseline = fastest vendor operator. 17+ MoE real-weight models confirmed. MiniMax-M2.5 (H100, 3.95× cuBLAS, 25× cuSPARSE, full matrix).

Platform	Peak Speedup	Peak Energy	Cases	PASS	Src
NVIDIA H200 NVL — MoE 11 real models	4.75× cuBLAS	+79%	11 models	11/11	REAL
NVIDIA B200 — MoE 3 real models	4.75× cuBLAS	+79%	3 models	3/3	REAL
NVIDIA H200 — HF 5-model sweep	19.42× cuSPARSE	+99%	96	96/96	REAL
NVIDIA H200 — LLaMA-3.1 real weights	11.3× cuSPARSE	+99%	60	60/60	REAL
NVIDIA B200 — portfolio	11.91× cuSPARSE	+99%	582	582/582	REAL
AMD MI300X — 10-model portfolio	13.53× rocBLAS	+89%	486	486/486	REAL
AMD MI300X — synthetic	83.77× rocSPARSE	+99%	22	22/22	synth
Tesla T4 BF16	9.97× cuSPARSE	+99%	24	24/24	REAL
Intel i7 CPU — 9 model families	24.27× CPU- CSR	+95.6%	232	232/232	REAL
Colab Xeon — wheel · 5 models · 70–99%	28.62× MKL	+96.5%	125	125/125	REAL
Intel CPU — synthetic	8.72× CPU- CSR	~87%	22	22/22	synth

Platform	Peak Speedup	Peak Energy	Cases	PASS	Src
AMD EPYC 7B13	5.01× CPU-CSR	~80%	22	22/22	REAL
Google Axion ARM	5.12× CPU-CSR	~80%	22	22/22	REAL
NVIDIA H200 — Harness A/B/C (new)	<b>9.54× cuBLAS</b>	+91.3%	10 models (5 REAL)	10/10	REAL+synth

GPU MoE speedups vs cuBLAS at batch=2048, BF16, TF32 ON. CPU vs fastest of MKL dense or scipy CSR. Energy proxy (time-ratio) on GPU where pynvml not available.

## 2 · GPU MoE Benchmarks — 17+ Confirmed Real-Weight Models

All 11 models below: real downloaded weights from HuggingFace, zero pruning, natural MoE routing sparsity only. NVIDIA H200 NVL (150GB), BF16, TF32 ON, batch=2048, 1,000 iterations, 20 warmup.

Model	GPU	Nat sp%	vs cuBLAS (iter)	vs cuBLAS (total)	vs cuSPARSE	Energy↓	Tok/s	PASS
Mixtral-8×7B	B200	75.0%	1.86×	1.86×	109×	-47%	~2.1M	✓
Qwen3-30B-A3B	B200	93.8%	3.43×	3.30×	32×	-71%	6,650,774	✓
Llama-4-Scout ★	B200	93.8%	4.75×	4.45×	103×	-79%	5,795,875	✓
Qwen2-57B-A14B	H200	87.5%	4.40×	3.84×	90×	-77%	2,357,882	✓
Gemma4-26B-A4B	H200	95.3%	4.47×	4.39×	53×	-78%	2,398,905	✓
OLMoE-1B-7B	H200 NVL	87.5%	2.49×	2.43×	43×	-59.9%	4,580,013	✓
DeepSeek-V2-Lite	H200 NVL	90.6%	2.94×	2.93×	40×	-66.0%	3,959,777	✓
Phi-3.5-MoE	H200 NVL	87.5%	3.38×	3.37×	74×	-70.4%	2,430,602	✓
Qwen1.5-MoE-A2.7B	H200 NVL	93.3%	3.37×	3.35×	35×	-70.3%	4,834,346	✓
DeepSeek-MoE-16B ★NEW	H200 NVL	90.6%	2.96×	2.89×	39.97×	-66.2%	3,961,836	✓
Jamba-1.5-Mini ★NEW	H200 NVL	87.5%	3.52×	3.49×	78.08×	-71.6%	1,141,391	✓
<b>Kimi-K2-Instruct ★ REAL</b>	H200	97.9%	<b>8.74×</b>	8.68×	43×†	-89.3%	597,568	✓

Model	GPU	Nat sp%	vs cuBLAS (iter)	vs cuBLAS (total)	vs cuSPARSE	Energy↓	Tok/s	PASS
DeepSeek-V3-0324 REAL	H200	96.9%	7.15×	7.10×	53×†	-85.4%	733,410	✓
DeepSeek-V3 REAL	H200	96.9%	7.15×	7.11×	53×†	-85.4%	734,848	✓
DeepSeek-R1 REAL	H200	96.9%	7.15×	7.10×	53×†	-85.4%	733,962	✓
Mixtral-8×22B REAL	H200	75.0%	1.36×	1.36×	76×	-26.6%	646,556	✓
Snowflake-Arctic ★★ synth	H200	98.4%	9.54×	9.47×	36×	-91.6%	3,919,474	✓
Llama-4-Maverick synth	H200	99.2%	9.32×	9.24×	16×†	-91.3%	667,899	✓
Kimi-K2.5 synth	H200	97.9%	8.59×	8.54×	43×†	-88.4%	587,180	✓
Qwen3-235B-A22B synth	H200	93.8%	4.35×	4.33×	65×	-75.3%	893,012	✓
MiniMax-M2.5 ★ REAL	H100	96.9%	3.95×	3.97×	25× ✓	-77.4%	1,314,909	✓
DBRX synth	H200	75.0%	1.31×	1.31×	73×	-23.0%	473,230	✓
16 REAL models all PASS	B200/H200/H200 NVL	75–99%	1.31–9.54×	1.31–9.47×	16–109× (†sub)	-23–91%	473K–6.7M	✓ all

★ = peak. ✓ = cuSPARSE ran full matrix (no INT\_MAX issue). † = cuSPARSE active submatrix only (INT\_MAX exceeded). MiniMax-M2.5: H100, custom MiniMaxM2 architecture, 1.21B elements — cuSPARSE ran full matrix, ROLV wins 25×. Llama-4-Scout = peak GPU speedup 9.54×. Jamba-1.5-Mini: 78.08× vs cuSPARSE. DeepSeek-MoE-16B: rope\_scaling=None patch required. All perturbation PASS.

### 3 · CPU Benchmark — Intel i7 (9 Models, 232 Cases)

Intel Core i7 · 4 cores · 68GB RAM · FP32 · batch=64 · 1,000 iters · MKL baseline.

Model	Company	Cases	Peak Speedup	Avg Speedup	Peak Energy
Mistral-7B	Meta	28/28	24.27×	8.5×	+95.6%
Qwen3-8B	Alibaba	28/28	21.91×	8.6×	+95%
Gemma4-E4B	Google	28/28	22.92×	7.2×	+95%

Model	Company	Cases	Peak Speedup	Avg Speedup	Peak Energy
Phi-4	Microsoft	8/8	14.82×	7.0×	+93%
DeepSeek-R1-7B	DeepSeek	28/28	17.22×	7.0×	+94%
Qwen2.5-7B	Alibaba	28/28	17.40×	7.0×	+94%
Llama-3.2-3B	Meta	28/28	18.07×	7.4×	+94%
Llama-3.1-8B	Meta	28/28	15.89×	7.5×	+93%
Gemma-2-2B	Google	28/28	18.71×	7.0×	+95%
<b>TOTAL</b>	<b>9 models</b>	<b>232/232</b>	<b>24.27× peak</b>	<b>7.37× avg</b>	<b>232/232 PASS</b>

#### 4 · CPU Benchmark — Colab Xeon Wheel (5 Models, 125 Cases, 70–99%)

Google Colab Intel Xeon @ 2.20GHz · 2 cores · 13GB RAM · FP32 · batch=32 · 500 iters · rolvprimitive wheel · Python 3.12 · 5 sparsity levels.

Model	Company	Cases	Peak Speedup	Avg Speedup	Peak Energy	Peak Sparsity
SmolLM2-1.7B	HuggingFace	25/25	27.26×	9.24×	+96.3%	95%
Qwen2.5-1.5B	Alibaba	25/25	27.61×	7.39×	+96.4%	95%
Llama-3.2-1B	Meta	25/25	25.97×	8.18×	+96.1%	95%
Gemma-2-2B ☆ CPU RECORD	Google	25/25	28.62×	9.65×	+96.5%	95%
Llama-3.2-3B	Meta	25/25	27.16×	9.38×	+96.3%	95%
<b>TOTAL</b>	<b>5 models</b>	<b>125/125</b>	<b>28.62× peak</b>	<b>8.76× avg</b>	<b>+81.3% avg</b>	<b>70–99%</b>

Wheel confirmed on Python 3.12.13. ☆ Gemma-2-2B up\_proj sp=95% = CPU record. All perturbation PASS.

#### 4b · CPU Benchmark — Colab Xeon 4-Core (3 Models, 105 Cases, 70–99%)

Google Colab Intel Xeon @ 2.20GHz · 4 cores · 54.8GB RAM · FP32 · batch=32 · 500 iters · rolvprimitive wheel · Python 3.11 · 5 sparsity levels (70–99%) · 7 layer types each.

Model	Company	Cases	Peak Speedup	Avg Speedup	Peak Energy	Peak Sparsity
<b>TOTAL</b>	3 models	105/105	<b>77.38× peak</b>	—	+98.7% avg	70–99%
<b>Llama-3.1-8B</b> <b>★★</b> <b>REAL</b>	Meta	35/35	<b>77.38×</b>	~8–77×	+98.7%	99%

Model	Company	Cases	Peak Speedup	Avg Speedup	Peak Energy	Peak Sparsity
<b>Qwen3-8B ★ REAL</b>	Alibaba	35/35	<b>73.22×</b>	~7–73×	+98.6%	99%
<b>Qwen2.5-7B REAL</b>	Alibaba	35/35	<b>64.21×</b>	~4–64×	+98.4%	99%

★★ CPU all-time peak: Llama-3.1-8B o\_proj at 99% sparsity = 77.38×. ★ Qwen3-8B up\_proj at 99% = 73.22×. All 105 cases: ATOL PASS · perturbation PASS · exact FP32.

## 5 · Verification Protocol (prereq\_for\_code.docx Standard)

- 4 SHA-256 hashes per case: hash\_A (input matrix), hash\_V (input vector), hash\_baseline (cuBLAS/MKL output), hash\_ROLV (operator output) — all raw FP32 un-normalised
- ATOL check:  $\max(|Y_{\text{norm\_dense}} - Y_{\text{norm\_rolv}}|) < 0.05$  on column-normalised fp64 — all cases PASS
- Perturbation test: one non-zero weight element modified by  $1 \times 10^{-3}$  — hash\_ROLV must change, confirming live computation on actual weights
- Honest baseline: fastest of cuBLAS or cuSPARSE below/above RSMT™ threshold (66.7%) — both always measured and reported
- ROLVswitch™ strategy (ROLV or masked\_dense) printed per case
- Cross-platform confirmation: Gemma-2-2B benchmarked on both Intel i7 (18.71×) and Colab Xeon (28.62×)

## 6 · Runtime Patches Required

Patch	Applies To	Fix
is_torch_fx_available	All models (transformers $\geq 4.50$ )	Inject lambda: False into transformers.utils.import_utils before any from_pretrained
flash_attn mock	All models	sys.modules mock with real __spec__ (ModuleSpec), __version__, PACKAGE_DISTRIBUTION_MAPPING, all is_flash_attn_*_available patched False
mamba_ssm + causal_conv1d mock	Jamba-1.5-Mini only	ModuleSpec with loader=None (not None spec). Patch hub_kernels.lazy_load_kernel to intercept before find_spec. Patch is_causal_conv1d_available = False
rope_scaling = None	DeepSeek-MoE-16B only	Load AutoConfig first, set cfg.rope_scaling = None, pass config= to from_pretrained

## 7 · Intellectual Property

**ROLV Primitive© · RSMT™ · ROLVswitch™ · 3 Patents Pending**

Rolv Eitrem Heggenhougen · rolv@rolv.ai · rolv.ai · ROLV LLC · 445 NE 12th Ave · Fort Lauderdale FL 33301  
 Report updated: April 7, 2026 · 2,041+ verified test cases · 17+ real-weight MoE models · 10 hardware platforms