Nemotron activation extraction

Output speed and numerical agreement — Antonio Lobo-Santos, DFKI — 21 October 2026

On one H200, the package saves activations at **5,535 tokens/s versus 3,930 for Transformers** at batch 64 (**1.41**×). The sibling fill writer reaches **24,079 tokens/s** (**6.13**×). We assess both output speed and numerical agreement of these activations: the model's internal vectors at selected recording points, or sites.

Output speed. We record 18 sites across six hook types in the official Nemotron 3 Nano BF16 checkpoint. For a consistent comparison, each table entry processes the same 131,072 frozen English tokens in 512-token sequences on the same physical H200, after two warm-up batches. Output tokens/s counts input tokens whose activations are saved at all sites, including capture, transfer, buffer waits and completed writes. Loading, warm-up and numerical comparison are excluded. GPU forward speed separately measures model execution.

Method	Output tokens/s Batch 8	Output tokens/s Batch 64	Output ratio Batch 64 / HF	Forward tokens/s Batch 64
Transformers (HF/PyTorch)	3,490	3,930	1.00×	24,679
Package vLLM asynchronous	6,406	5,535	1.41×	65,481
Sibling vLLM file sink	2,986	2,724	0.69×	67,474
Sibling vLLM fill writer	20,928	24,079	6.13×	67,087

The package peaks at 6,406 output tokens/s across five batch sizes (8–128): 1.84× Transformers at batch 8. At batch 64, however, similar vLLM forward speeds yield markedly different output rates. Model execution speed alone therefore does not predict extraction throughput. These complete paths differ in publication and finite-value checks. All use eager BF16 execution; vLLM enables chunked prefill and automatic KV-cache sizing. Requested batch sizes limit admission; actual forward batches may be smaller.

Numerical agreement. To assess output agreement, we compare each vLLM path with Transformers (HF/PyTorch). Cosine measures directional alignment; relative L2 measures error relative to the reference magnitude. Each site must achieve mean cosine ≥ 0.9999, minimum token cosine ≥ 0.9990, and relative L2 ≤ 0.02. All three paths pass only 3/18 sites. Similarly, a BF16 sample of 16,384 tokens per language gives mean site cosine 0.999002 (English), 0.998623 (German) and 0.998701 (Spanish). High averages therefore conceal token outliers. Normalised MSE equals relative L2 squared.

Explaining the differences. To separate local arithmetic error from upstream differences, we replay individual operations with original weights and identical saved 2,048-token inputs. On RTX A6000, PyTorch and vLLM mixture-of-experts (MoE) operations share a runtime, FP32 routing, top-six selection among 128 experts and normalised weights scaled by 2.5. Both use IEEE settings, disable reduced BF16 reductions and share the selector to align ties. Across layers 1, 27 and 51, mean cosine rises from 0.9999948–0.9999995 in BF16 to 1.0000000 in FP32 (rounded); worst-token cosine rises from 0.9999882 to 0.9999993. Maximum relative L2 falls from 0.00365 to 1.05×10^{-6}; all 16 comparisons pass. Likewise, the first Mamba block on RTX 3090 gives relative L2 0.00291 in BF16 and 1.18×10^{-7} in FP32. IEEE flags alone leave its BF16 output unchanged.

We then assess expert selection using the differing saved backend inputs. Keeping vLLM's input fixed, we force the reference expert IDs and recompute their weights in BF16. Each layer has 2,048 tokens:

Layer	Changed sets	Single-expert swaps	Tokens with cosine < 0.99	Squared error removed
1	39	39	$4 \rightarrow 0$	64.6%
27	356	342	$276 \rightarrow 73$	40.5%
51	261	256	$7 \rightarrow 3$	17.4%

At cosine < 0.999, token counts fall from 28 to 0, 1,005 to 885 and 139 to 96, respectively. Among tokens recovered above 0.99, single-expert swaps account for 4, 193 and 3 recoveries. Defining a near tie as a corrected sixth/seventh score gap ≤ 0.001 gives only 3, 23 and 0 recoveries. Selection therefore explains some discrepancies, particularly at layer 1, while deeper errors extend beyond near ties. These effects interact with upstream error and cannot be added across layers.

Scope and evidence. The output rates describe package 1.0.0 in BF16; the FP32 results validate individual operations. Full-model FP32 agreement and output speed remain unmeasured. All runs pin the same checkpoint revision. Source hashes and timings: `docs/extraction_speed_results.json`. Arithmetic configuration and evidence: `docs/matched_precision.json`, `docs/moe_validation.md`.