Local Execution and Hybrid Topology Checkpoint for NVIDIA-Nemotron-3-Super-120B-A12B-BF16

VertRule technical note

2026-04-11 · v2 · NVIDIA-Nemotron-3-Super-120B-A12B-BF16 · CPU

1. Claim

NVIDIA-Nemotron-3-Super-120B-A12B-BF16 executes locally under governed streamed execution with a frozen anchor, canonical topology emission, determinism-gated admission, and a bounded hybrid-edge surface combining expert-routing and attention edges.

The model is a 120-billion-parameter hybrid architecture with 88 layers: 40 linear-attention (Mamba2 SSM), 40 mixture-of-experts, and 8 full-attention. Each MoE layer routes tokens to 22 of 512 experts via a sigmoid-scored top-k routing surface.

This checkpoint shows that the local governed execution and topology surface extends to a substantially larger hybrid model with three distinct layer types.

2. Contract

All results below were verified on a local CPU-only admitted measurement surface using real BF16 weights (230 GB, 50 safetensors shards, 42,683 tensors) streamed from local storage on an Apple M4 Pro.

- Bounded setups. Single-token forward (frozen non-perturbation baseline) and two-token sequential prefill (attention-edge capture).
- Execution mode. Per-layer weight eviction, no GPU or accelerator involvement.
- Determinism. Bit-exact across multiple independent forward variants.
- Capture non-perturbation. The frozen digest is preserved under full (88-layer), empty (0-layer), and sparse ({0, 44, 87}) capture configurations.
- Topology reproducibility. Canonical topology digests are identical across two independent capture-and-assembly runs.
- Determinism gate. All structural invariants pass on every admitted topology phase (zero-edge, expert-routing-edge, and mixed-edge).

3. Runtime Anchors

Three structural anchors pin the model's identity. These have been verified across multiple independent read paths and are unchanged across all phases of this checkpoint.

A frozen anchor pins the single-token forward-pass output. The current canonical anchor supersedes a historical anchor from an earlier implementation epoch. The historical anchor remains valid for that earlier code era.

The frozen digest is validated across multiple independent execution variants. All produce bit-identical results.

4. Topology Anchors

A canonical `topology.v2` artifact is emitted through a shared topology assembly surface.

Three topology phases have been admitted for this checkpoint:

- Zero-edge phase. 88 nodes (one per layer), 0 edges, concept scores from a synthetic 4-concept bank. This phase established the baseline canonical topology and was admitted through the topology determinism gate.
- Expert-routing-edge phase. 88 nodes, 880 edges (40 MoE layers $\times$ 22 experts). Re-digested and re-admitted through the same gate.
- Mixed-edge phase. 88 nodes, 896 edges (880 expert-routing + 16 attention). The attention edges are captured from a two-token sequential prefill at query position 1 over the 8 full-attention layers (top-2 head-averaged per layer). Re-admitted through the same gate.

All three phases produce reproducible canonical digests across independent runs. All pass all determinism gate invariants.

5. Bounded Hybrid Edge Surface

The bounded edge surface combines two edge kinds:

Expert-routing edges (880). Each of the 40 MoE layers selects 22 experts per token from a pool of 512. For each selected expert, an edge is emitted with:

- src: token position
- dst: expert index (0–511)
- kind: expert routing
- weight: the normalized-and-scaled routing weight used in the actual weighted expert combination

Attention edges (16). Each of the 8 full-attention layers produces a head-averaged attention distribution at query position 1 over the two cached key positions from the two-token prefill. The top-2 keys by probability (descending, key-position-ascending tie-break) are emitted as attention edges.

The total bounded surface is 896 edges (880 expert-routing + 16 attention), deterministically ordered by layer index, source, and destination for canonical gate compliance.

Expert routing is the dominant emitted edge kind (880 of 896 edges) and directly expresses the model's sparse activation structure. Mamba2 layers contribute topology nodes but no edges; the SSM state is internal to the mixer and not visible at the residual-stream layer boundary.

Bounded observation

For the canonical input token 1784 on Nemotron-3-Super-120B, the median expert-routing concentration ratio of the later 20 MoE layers is 4.02, versus 1.51 for the earlier 20 MoE layers —

a 2.7× split — where concentration ratio is the normalized-and-scaled routing weight of the highest-weighted selected expert divided by that of the lowest-weighted selected expert within each layer. The common scale factor cancels in the ratio.

The observation is computed from the already-admitted expert-routing topology surface whose 880 edges are bit-identical across independent runs.

The partition is structural: the 40 MoE layers are split into the earlier 20 and later 20 MoE layers, not at a threshold selected to maximize the ratio. The metric ($w_{\max}/w_{\min}$ among the 22 selected experts) is the natural concentration summary for a fixed top-22 router. The median is used rather than the mean to resist the single early outlier at layer 5 (ratio 5.41).

This observation does not claim:

- That the later-half concentration increase reflects expert specialization, topic separation, or any semantic property
- That the depth split is monotonic (it is not — individual layers deviate from the aggregate trend)
- That the 2.7× split generalizes to other input tokens, prompts, or model variants
- That the concentration ratio has a causal relationship to model quality or task performance

6. Boundary

This checkpoint does not claim:

- A bounded Mamba state-transition surface
- Browser, Metal, or cross-backend execution
- Cross-platform reproducibility beyond a single M4 Pro
- Multi-token or autoregressive determinism (only single-token is anchor-gated)
- Long-context behavior
- Semantic or mechanistic conclusions from the expert-routing surface
- Final hybrid-topology completeness — the current mixed-edge surface covers expert-routing and attention edges, but not bounded Mamba state-transition edges

7. Verification

Property	Status
Forward-pass determinism across multiple independent variants	Verified
Capture non-perturbation (full / empty / sparse)	Verified
Canonical topology reproducible across independent runs	Verified
Determinism gate pass — zero-edge phase	Verified
Determinism gate pass — expert-routing-edge phase	Verified
Determinism gate pass — mixed-edge phase	Verified
Frozen single-token digest preserved at step 0 of two-token prefill	Verified
Two-token step-1 logits deterministic across independent runs	Verified
Frozen digest unchanged under topology emission	Verified
Mixed-edge count = 896 (880 expert + 16 attention)	Verified
Node count = 88	Verified
Schema version = `topology.v2`	Verified

Anchor table

All digests and receipts below are BLAKE3.

Anchor	Value	Size
Config digest	`c515c1614451103b7dc71ab236ec693e291bf34d939aad2a38414f46d03f4223`	
Tensor-inventory digest	`a76bb291413f595ffc72d133f7a02acb0b246c5431571465d391214fa868d1d83`	41,683 tensors
Operator-graph digest	`576f2d7e56c0741f52eae5fcb8cfcc4981638515e52b67f427d9ef3eea1e07a6`	17,639 nodes
Frozen digest (canonical)	`4f230ac48c05b3e591f55337354a49376b22980f7428b11d4a6c4d4991750cbd`	
Historical digest (pre-transition)	`d7fd56a666fba4930ca0b3bff7aa0f19ce85a942c4ecf59bd04aefb550aa79f0`	
Zero-edge topology digest	`9e370370363e685798e5043c3b708ac1a2103268ba8e5b8c8bb6b790938988ec7`	88 nodes, 0 edges
Expert-edge topology digest	`ccfac095a399d5552e9ec0d7c2bade1dd4331bb0fd9f528a16e7639b50a788e9`	88 nodes, 880 edges
Zero-edge determinism gate receipt	`ef6c32990da6423bb28a483cbd8b3bc79311b4d64a8b94862cbda8d53135e762`	
Expert-edge determinism gate receipt	`d827fe9e603ae2f6fd12e400ec54e054c81e5464055a0deb645680deb39a2b43`	
Two-token step-1 logits digest	`b40639bdaa51b95f94454a04a1fb6c11bb9d059f2116868380fe4aae4038f2b1`	
Mixed-edge topology digest	`bda232875ac85cedace62938b16813b8a06b8191e5431a99ae06400b4edb8710`	88 nodes, 896 edges
Mixed-edge determinism gate receipt	`08ef945a0fefa9a2decdc7513520cb4dadf31869501f1a8f0bb96902da0f02b1`	

8. What Remains

This note records the current admitted mixed-edge topology surface. Future work may extend it with a bounded Mamba state-transition edge surface.