NAS-Backed Local Execution and Bounded Topology Checkpoint for Grok-1 314B

VertRule technical note

2026-04-12 · v1 · Grok-1 314B · NAS-backed CPU

1. Claim

Grok-1 314B executes locally on a NAS-backed admitted measurement surface with a ratified runtime anchor, non-perturbing capture, canonical topology admission, and a bounded mixed-edge surface combining expert-routing and attention edges.

The model is a 314-billion-parameter mixture-of-experts architecture with 64 layers. Every layer contains full grouped-query attention (48 query heads, 8 key-value heads, head dimension 128) followed by a top-2-of-8 expert routing surface with SwiGLU expert MLPs. The vocabulary size is 131,072. Weights are INT8-quantized with BF16 per-group scales.

This checkpoint shows that the local governed execution and topology surface extends to a model stored entirely on a network-attached volume accessed over SMB, where the storage contract is part of the admitted measurement surface.

2. Contract

All results below were verified on a local CPU-only admitted measurement surface using real INT8+BF16 weights (296 GB, 770 safetensors shards) read from a network-attached storage volume over SMB.

- Storage contract. Weights reside on a NAS volume mounted via SMB. The runtime anchor was ratified under a page-cache-purge protocol: the operating system unified buffer cache was purged between each ratification run, forcing cold reads from the network volume at each pass. Four independent purge-separated runs (wall times 43.81 s, 43.93 s, 43.98 s, 44.01 s) produced identical digests, consistent with cold sequential NAS throughput of approximately 97 MB/s over 4.3 GB of weight data per forward. This contract does not claim that all server-side or client-side caching is eliminated globally; only that the named purge-based protocol was applied and that the digest was stable under it.
- Bounded setups. Single-token forward (1 token, 1 layer; frozen non-perturbation baseline) and two-token sequential prefill (2 tokens, 2 layers; mixed-edge capture).
- Execution mode. Per-layer expert deduplication (each unique expert loaded once from NAS per layer, evicted at layer boundary), CPU-only, release profile, aarch64-apple-darwin. No GPU or accelerator involvement.
- Determinism. Bit-exact across independent runs under the named storage contract.
- Capture non-perturbation. The frozen digest is preserved under full (all layers), empty (no layers), and sparse (single layer) capture configurations for both layer-boundary and attention-weight capture.

- Topology reproducibility. Canonical topology digests are identical across two independent capture-and-assembly runs at every admitted phase.
- Determinism gate. All structural invariants pass on every admitted topology phase (zero-edge, expert-routing-edge, two-layer expert-routing-edge, and mixed-edge).

3. Runtime Anchors

Two structural anchors pin the model's weight identity:

- A config digest over the canonical model geometry (hidden size, head counts, expert counts, vocabulary size, norm epsilon, RoPE theta).
- A tensor-inventory digest over the sorted file-size listing of all 770 weight shards.

No on-disk configuration file exists for this checkpoint. The model geometry is encoded as constants derived from the original release.

A frozen anchor pins the single-token forward-pass output. The anchor was ratified under the page-cache-purge protocol described in the contract. The anchor scope is deliberately narrow: one input token (vocabulary index 1), one MoE layer, CPU-only, release profile. This is the smallest honest admitted surface; wider scopes can be anchored independently.

4. Topology Anchors

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

Four topology phases have been admitted for this checkpoint:

- Zero-edge phase: 1 node (1 layer, 1 step), 0 edges, concept scores from a synthetic 4-concept bank. This phase established the baseline canonical topology for the NAS-backed execution path and was admitted through the topology determinism gate.
- Expert-routing-edge phase: 1 node, 2 edges (1 MoE layer, top-2 experts). Re-digested and re-admitted through the same gate.
- Two-layer expert-routing phase: 2 nodes, 4 edges (2 MoE layers, top-2 experts each). Admitted through the same gate.
- Mixed-edge phase: 4 nodes, 8 edges (4 expert-routing + 4 attention). The attention edges are captured from a two-token sequential prefill at query position 1 over 2 layers (top-2 head-averaged per layer). Admitted through the same gate.

All four 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 (4). Each of the 2 admitted MoE layers selects 2 experts per token from a pool of 8 via softmax-normalized top-k routing. For each selected expert, an edge is emitted with:

- src: token position

- dst: expert index (0–7)
- kind: expert routing
- weight: the renormalized routing weight used in the actual weighted expert combination

Attention edges (4). Each of the 2 admitted 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 8 edges (4 expert-routing + 4 attention), deterministically ordered by layer index, source, and destination for canonical gate compliance.

Expert routing and attention edges are equally represented in this bounded surface, reflecting Grok-1's architecture where every layer contains both full attention and MoE routing.

Bounded observation

For the canonical input token 1 on Grok-1 314B, the expert-routing concentration ratio decreases monotonically over the first 4 layers, where concentration ratio is the renormalized routing weight of the highest-weighted selected expert divided by that of the lowest-weighted selected expert within each layer:

Layer	Selected experts	Weights	Concentration
0	5, 1	0.7238, 0.2762	2.62
1	0, 3	0.6830, 0.3170	2.15
2	0, 1	0.5179, 0.4821	1.07
3	3, 2	0.5138, 0.4862	1.06

The router selects a strongly dominant expert at layer 0 (concentration 2.62) and converges toward near-equal weighting by layer 3 (concentration 1.06). No expert index appears in more than two consecutive layers.

This direction of change is opposite to what has been observed on a different architecture (Nemotron-3-Super-120B, where concentration increases with depth from 1.51 to 4.02 over 40 MoE layers). The two models differ in pool size (8 vs 512), selection width (top-2 vs top-22), and routing mechanism; the contrast is structural, not interpretive.

Corroborating this direction, head-averaged attention concentration at query position 1 on the two-token prefill also decreases with depth over the admitted 2-layer mixed-edge surface:

Layer	Top-2 key probs	Attention concentration
0	0.8933, 0.1067	8.37
1	0.5670, 0.4330	1.31

The attention observation is computed from the admitted mixed-edge topology surface. The routing observation is verified as bit-identical across independent runs over the 4-layer observation surface.

This observation does not claim:

- That the concentration decrease reflects expert specialization, diffusion of knowledge, or any semantic property
- That the monotonic decrease continues beyond 4 layers (only 4 of 64 layers are in the observation surface)
- That the concentration ratio generalizes to other input tokens, prompts, or model variants
- That the directional contrast with other architectures has a causal explanation
- That the concentration ratio has a relationship to model quality or task performance

6. Boundary

This checkpoint does not claim:

- Browser, Metal, or cross-backend execution
- Cross-platform reproducibility beyond aarch64-apple-darwin
- Execution parity with local-resident storage (only the NAS-backed storage contract is admitted)
- Multi-token or autoregressive determinism beyond the admitted bounded surfaces (1-token/1-layer anchor; 2-token/2-layer mixed-edge)
- Behavior beyond the first 4 layers of a 64-layer model
- Long-context behavior
- Semantic or mechanistic conclusions from the expert-routing or attention surfaces
- Full-model inference or serving capability
- That the page-cache-purge protocol eliminates all forms of network-layer or server-side caching; only that the named protocol was applied and the digest was stable under it

7. Verification

Property	Status
Forward-pass determinism across purge-separated NAS reads	Verified
Capture non-perturbation (full / empty / sparse)	Verified
Attention capture non-perturbation (2-token prefill)	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 – two-layer expert-routing phase	Verified
Determinism gate pass – mixed-edge phase	Verified
Mixed-edge count = 8 (4 expert + 4 attention)	Verified
Node count = 4 (mixed-edge phase)	Verified
Schema version = `topology.v2`	Verified
Expert concentration monotonically decreasing over 4 layers	Verified

Property	Status
Input-dependent routing (token A vs token B)	Verified
State contamination (A then B then A: second A matches first)	Verified
Gate falsifier (poisoned digest triggers rejection)	Verified
NAS cache-busting (15x wall-time ratio cold vs warm)	Verified

Adversarial verification

Four adversarial tests verify that the runtime is dynamically tracking the model's physical state and that the verifier is not blind.

Input-dependent routing. A second input token (vocabulary index 9999) produces the same expert selection at layer 0 but with different renormalized routing weights (0.7000, 0.3000 vs 0.7238, 0.2762) and a different logits digest. The observation lane is tracking dynamic activation state, not returning hardcoded values.

State contamination. An A-then-B-then-A sequence (token 1, then token 9999, then token 1 again) produces a second token-1 run that is bit-identical to the first: identical digest, identical expert indices, identical routing weights. No activation state bleeds across runs.

Gate falsifier. Flipping a single hexadecimal character in the observed logits digest causes the determinism gate to immediately reject with two violations: digest drift from the frozen anchor and digest non-reproducibility between runs. The gate is not blind.

NAS cache-busting. Purge-separated cold runs complete in 43.81–44.01 seconds. Warm runs (page cache hot) complete in 2.4–3.0 seconds. The 15x wall-time ratio confirms the purge protocol forces genuine NAS reads.

Anchor table

All digests and receipts are BLAKE3.

Anchor	Value	Size
Config digest	`458f34857b4a61dd428a1afea90ab00a136a79180edf585c6941606b51dee412`	
Tensor-inventory digest	`c3e3299a6b0fe8b601c151dd9d87d3d7a6137ec6ef94624453f541cc5b21578d`	5783 shards
Frozen digest (1-token, 1-layer)	`f7356e44dcccd5b876a525ea91b73f62160578fd595fb00c9f46d0908653bace`	
Zero-edge topology digest	`4005dd2e3069531dc15a561358056a24e13acdb673f3e41b8780beb439c418dd`	1 node, 0 edges
Expert-edge topology digest	`c35994b40c4d3a893391792c557d72ac9b8d6e45e1acaf4ed43585552c08b26b`	4 nodes, 2 edges
Two-layer expert topology digest	`c034345ef5c378185e4dda030adbb01bbcc55af7be018b479e76db2a64d42bbd`	2 nodes, 4 edges
Mixed-edge topology digest	`873a34e26667e13f37c96c7d0f533fe30b06bed6215a58084edc78dcd9dd1e60`	4 nodes, 8 edges
Zero-edge determinism gate receipt	`7da33ff4f647988915fc93b6b252dd7372ed56116bddf34a4ba64f01d0cc0126`	
Expert-edge determinism gate receipt	`e51a23cd9f2d15edae1395b2b609e29b83bd496a328aeffd90dbcdb24119536 9`	
Two-layer expert gate receipt	`83d522d0470d28427610df27513eaa71dc033e27f2de624f80718284521a0569`	
Mixed-edge determinism gate receipt	`a4af9b7bafbeaf8e34ca8393d1973c40c35e91b80103ec7373166a427b07c5c2`	

8. What Remains

This note records the current admitted NAS-backed mixed-edge topology surface over the first 4 layers of a 64-layer model. Future work may extend the admitted surface to deeper layers, a broader token set, or a bounded autoregressive continuation surface. The operator-graph digest for the full 64-layer model has not yet been computed.