Full-Weight Local Execution and Output Formation Ledger for Kimi K2.5

VertRule technical note

2026-04-13 · v1 · Kimi K2.5 · CPU + NAS

Kimi K2.5 — a 595 GB, 61-layer DeepSeek-V3 model with 384 MoE experts — completes a full-weight, full-layer, NAS-backed local inference pass on a consumer laptop. Five tokens are generated autoregressively, each traversing all 61 transformer layers with weights streamed from network storage.

The model's output formation is exposed through a token-by-token execution ledger: 305 layer observations across 5 generation steps, with every surfaced value tagged as measured or derived under a provenance contract. The execution surface is frozen as a deterministic anchor, the capture instrument is proven non-perturbing, and the full artifact set has passed adversarial certification.

`M+D · Lab · Ratified · LC+NS · X0 · NP · 5 tok / 61 layers / no-cache full pass · Output formation ledger`

This is not a benchmark. It is not a partial forward pass. It is not a topology checkpoint. It is a complete, honest, inspectable proof that a 595 GB model can execute locally and that its output formation can be observed without altering it.

---

1. Claim

Kimi K2.5 completes a full streamed local inference pass on a consumer laptop, and its output formation is exposed through a token-by-token execution ledger that can be inspected under an explicit provenance and execution contract.

This is not a benchmark, a partial forward pass, or a claim about model reasoning or intention. The ledger captures output formation mechanics — which experts were selected, how residual norms evolved, what the logit distribution looked like — not why the model produced its output in any semantic sense.

2. Contract

Model. Kimi K2.5 is a vision-language model built on the DeepSeek-V3 text backbone. Only the text path is executed; vision components are not loaded.

Table 1. Model summary.

Property	Value
Architecture	DeepSeek-V3
Layers	61 (1 dense + 60 MoE)
Hidden dimension	7168
Attention	MLA, 64 heads, q_lora_rank=1536, kv_lora_rank=512
RoPE	YaRN interleaved, theta=50000
MoE	384 routed experts, top-8, sigmoid routing with correction bias
Shared expert	1 per MoE layer, always active, BF16
Expert quantization	INT4 symmetric group (group_size=32)
Non-expert weights	BF16
Vocabulary	163,840 tokens
Checkpoint	~595 GB total (~554 GiB), 64 safetensors shards

Execution properties.

- Local CPU compute (Apple Silicon). No GPU, no remote compute.
- NAS-backed weight storage, mounted over local network.
- Weights streamed per-layer from NAS-backed storage. Only activated tensors are read per step.
- No application-level KV cache. Full sequence recomputed per token step.
- 8 of 384 experts loaded per token per MoE layer. Each forward pass reads ~6% of the full checkpoint.
- OS page cache present after the first token step. Explicitly declared, not hidden.

Per-token I/O ranges from ~33 GiB (single token, 8 unique experts per MoE layer) to ~64 GiB (5-token context, up to 40 unique experts per MoE layer after deduplication across tokens). The full checkpoint is ~595 GB, but MoE selectivity means each pass loads only the activated expert subset.

Trust tier. The execution surface is at Laboratory trust tier. Proof-surface contracts governing provenance, synthetic auditing, perturbation verification, and anchor lifecycle are governed at a stricter trust tier. Ratification applies to the frozen execution anchor, not automatically to every derived field in the ledger.

3. Frozen Anchor and Determinism

The execution surface is anchored by a single-token proof: BOS token in, all 61 layers traversed, deterministic argmax token out. The anchor was established by running this path twice and verifying bitwise-identical results.

Table 2. Frozen single-token anchor.

Anchor	Value
Output token	950
Output logit	27.711794

Anchor	Value
Bytes loaded	35,335,326,976

Frozen digests (BLAKE3).

Hidden state:
`b3751954b18beb183a9d5f7b0601a31b28d74e36b573e8104c4007c3b970367a`

Logits:
`d24500ce107223c9a859c723211b59d43d96491a77459d3b97403d036f6f1341`

Output token, digests, and byte count are fixed reference values. The output logit is covered by the logits digest. Every subsequent run — instrumented or multi-token — asserts exact match on the first token step.

4. Non-Perturbing Capture

The output formation ledger requires per-layer measurements (residual norms, routing decisions) added to the forward pass. The instrument was proven non-perturbing: running with capture enabled produces identical output token, hidden digest, logits digest, and byte count to the frozen uninstrumented anchor.

This is the load-bearing trust condition. If any value had differed, the capture data would be untrustworthy and the ledger would not have been produced.

5. Output Formation Ledger

Each generated token produces a ledger entry containing 61 layer observations. Every field is classified by provenance:

Measured (M) — direct from execution:

- Output token ID and logit value
- Per-layer residual L2 norm before and after each layer block
- Per-layer MoE expert indices (which 8 of 384 were selected)
- Per-layer bytes read from storage

Derived (D) — deterministically computed from measured data:

- Top-16 tokens by logit value (sorted extraction)
- Shannon entropy of softmax logit distribution
- Margin between top-1 and top-2 logit values
- Normalized routing weights (sigmoid scores of selected experts, sum-normalized)
- BLAKE3 digests of hidden state and logits tensors

No interpretive content is present. The classification is structural — determined by the schema definition, not by runtime annotation.

6. Multi-Token Continuity

Five tokens were generated autoregressively from a BOS prompt.

Table 3. Multi-token execution summary.

Step	Context	Token	Logit	Bytes	Wall time
0	1	950	27.71	33 GiB	426.7s
1	2	1268	12.38	44 GiB	195.8s
2	3	426	17.38	52 GiB	158.6s
3	4	120287	15.42	59 GiB	128.7s
4	5	63	18.11	64 GiB	108.0s

Total: 251 GiB streamed, 1017.8 seconds wall time. Exact byte counts in verification summary.

Each step's input context contains all prior generated tokens. The first step preserves the frozen single-token anchor exactly. Per-step bytes grow because longer contexts activate more unique experts across MoE layers.

The timing gradient (427s to 108s) reflects OS page cache warming, not algorithmic improvement. The first step performs cold NAS I/O; subsequent steps find most shard data in RAM. These numbers describe one specific run under one specific infrastructure configuration and are not throughput claims.

7. Observed Structural Surfaces

The ledger reveals structural patterns in execution state. These are measured surfaces (M) unless noted. Phase labels are descriptive convenience, not model-endorsed categories.

Residual norm trajectory. Across 61 layers, the L2 norm of the hidden state follows a characteristic multi-phase trajectory visible in the explorer's norm chart. This shape is consistent across all five token steps. Phase labels in the explorer are visual summaries over measured norms, not a canonical segmentation of the model.

Expert routing diversity. Each MoE layer selects 8 of 384 experts per token. Selected sets vary across layers, indicating broad utilization of the expert pool. Routing weights (D, normalized) show non-uniform distributions.

Dense-to-MoE transition. Layer 0 uses a dense MLP (18432 intermediate, BF16) rather than MoE (384 experts, 2048 intermediate, INT4). This structural boundary is visible in byte accounting and routing data.

8. Verification

Adversarial certification. The full artifact set was subjected to hostile review.

Table 4. Adversarial certification results.

Check	Result
Anchor continuity (single / multi-token)	Verified
Cross-token input sequences	Verified
Synthetic surface audit	X0 (clean)
Provenance labels (305 layer entries)	All correct

Check	Result
Explorer scope	No inflation
Runtime contract	Stated honestly
Ratification scope	Correctly scoped to execution surface

Verdict: **SAFE** — no synthetic dependencies in the final artifact set.

Explorer. An interactive explorer renders the ledger as a navigable surface: token timeline, per-layer norm trajectory chart, layer detail table, and expert routing panel. Every displayed value carries a provenance indicator (green for measured, blue for derived). No interpretive content is rendered. The title is "Kimi K2.5 Output Formation Ledger."

Proof-surface contracts. Proof-surface contracts governing provenance classification, synthetic auditing, perturbation verification, and anchor lifecycle have been independently formalized.

9. Boundary

This work does not claim:

- That the generated token sequence is meaningful text. The output is raw autoregressive continuation from a BOS prompt without a chat template.
- That the ledger captures "reasoning." Norms, routing decisions, and logit distributions are execution telemetry, not evidence of semantic understanding.
- Speed or throughput. Five tokens in 17 minutes is a proof of complete, honest, inspectable local execution.
- Generalization of the NAS storage contract. Timing and stability reflect one specific infrastructure configuration.
- That every derived value is ratified. Ratification applies to the frozen execution anchor. Entropy, margin, and routing weight normalization are correctly computed but not independently certified as the only valid derivation.
- Semantic meaning in norm trajectory phases. The growth-ascent-plateau-decay-spike pattern is a measured structural observation, not a model-endorsed boundary.

What this work does establish: a 595 GB model can execute locally, completely, and honestly, and its output formation can be observed without altering it — under a provenance contract strict enough to survive adversarial review.

`M+D · Lab · Ratified · LC+NS · X0 · NP · 5 tok / 61 layers / no-cache full pass · Output formation ledger`