The Mask Is Not the Model: Auditing Prefix Invariance in Attention, State-Space, and Hybrid Sequence Models

Taebong Kim **Youngsik Hong** **Minsik Kim** **Sunyoung Choi**
Jaewon Jang Minseo Kim
VIDRAFT AI Research · QuantumOS, Seoul, Republic of Korea
arxigpt@gmail.com

Abstract

We formalize prefix invariance: representations at position t must not depend on future inputs. We give a lightweight audit, two forward passes, no training or gradients, that localizes exactly where causality breaks. Attention-mask inspection is incomplete: leaks can occur via scans or normalization despite correct masks. Across 192 injected-fault trials on eight checkpoints, mask inspection found none, while our audit localized all 192/192, also finding a defect in Zamba2 and Nemotron-H.

Keywords: Causal Leakage Detection,Prefix Invariance,Layer-wise Localization,Hybrid Language Models,Model Auditing

1 Introduction

1.1 A property nobody checks

Autoregressive language models have become the dominant paradigm for modern sequence modeling and underpin systems ranging from GPT-style language models to instruction-tuned assistants [1, 2, 3].

More broadly, modern foundation models emerged from a long line of sequence-learning research, including sequence-to-sequence learning [4], attention-based neural machine translation [5], bidirectional transformer pretraining [6], and generalized autoregressive objectives [7].

An autoregressive model is meaningful only if it satisfies one structural constraint: the output at position t may depend on inputs at positions $\leq t$, and on nothing else.

Every practical use of such a model—teacher-forced training, perplexity evaluation, incremental decoding, KV-cache reuse, and speculative decoding—silently assumes this property. Yet contemporary model releases routinely report parameter counts, context lengths, training tokens, and benchmark performance while providing no evidence that the released implementation is causally correct.

Historically this omission was less concerning because architectures were relatively homogeneous. Decoder-only Transformers contain a single dominant temporal mixing mechanism, self-attention, whose causal behavior is governed by an explicit mask.

The modern self-attention paradigm originates with neural attention mechanisms [5] and was standardized by the Transformer architecture [8].

Consequently, practitioners often treated inspection of the mask as a de facto audit of causality.

This assumption no longer holds. Recent sequence-model architectures increasingly combine heterogeneous mixers within a single stack, including local attention [9], sparse attention [10, 11], memory-augmented attention [12], linear recurrences [13, 14], structured state-space models [15, 16], state-space dual architectures [17], recurrent-attention hybrids [18, 19], and convolutional alternatives [20].

More broadly, efficient-sequence-modeling research now spans sparse, local, recurrent, convolutional, and state-space approaches designed to overcome the quadratic cost of standard self-attention [21].

In such systems, explicit attention masks govern only a subset of layers and may be absent altogether. The traditional audit procedure therefore stopped covering most of the computation graph.

1.2 Why silent leakage is the dangerous failure mode

Causal leakage is not a crash, an exception, or an obvious malfunction. Instead, it is a failure mode that can make development metrics appear *better*.

If future information reaches earlier positions, the next-token prediction task becomes artificially easier. Training loss decreases, validation perplexity decreases, and benchmark scores computed from teacher-forced likelihoods may improve, despite the implementation violating the assumptions of autoregressive inference. Similar forms of leakage have repeatedly produced misleading conclusions in machine learning research and have been identified as a major contributor to reproducibility failures across scientific domains [22].

The practical danger is asymmetry: leakage improves the very metrics used to select models while only becoming apparent during free-running generation, deployment, or downstream evaluation. By the time the issue is discovered, architecture comparisons and performance claims may already have been contaminated. This observation motivates a structural correctness check that is simple enough to run routinely and inexpensive enough to become part of standard evaluation practice.

1.3 Motivation: three faults in a row

This work began as an engineering tool rather than a research project. During development of an internal hybrid sequence-model stack, we encountered three distinct causal failures:

1. **A doubled output shift.** Two independently correct alignment operations combined to expose one future token.
2. **An incorrect causal condition in a custom attention implementation.** A hand-written attention path replaced a causal neighborhood with a symmetric one.
3. **A differential-attention aggregation fault.** A subtraction between attention maps was performed after a sequence-wide aggregation step had already mixed information across time.

Each fault required substantial manual debugging despite violating the same underlying property. None produced obvious anomalies in conventional training telemetry. The recurring pattern suggested that causal correctness should not require bespoke reasoning about individual architectures. Instead, it should be testable through a uniform property of the forward computation itself.

This perspective is closely related to the philosophy of metamorphic testing, where correctness is verified through invariance relations rather than task-specific labels [23]. Similar ideas have successfully exposed defects in deep-learning systems through automated testing frameworks such as DeepXplore and DeepTest [24, 25]. Our position is that autoregressive language models admit a particularly natural metamorphic relation: agreement of all prefix representations under perturbations that occur strictly in the future. Unlike conventional testing settings, the oracle is intrinsic to the definition of autoregressive causality itself.

The broader motivation is not confined to language models. Sequence modeling now spans language, speech, vision, and time-series domains, including WaveNet-style autoregressive audio generation [26], Vision Transformers [27], and deep time-series models [28]. A general-purpose correctness audit therefore has potential value beyond a single architecture family.

1.4 Contributions

Our contributions are:

1. **A named property and a test.** We formalize *prefix invariance* and introduce a two-forward-pass, gradient-free audit that identifies not only whether a causality violation exists but also the first layer where it appears.
2. **An incompleteness argument.** We show that mask inspection is not a sound test for causal correctness because causality is a property of the entire computational graph rather than of any single operator.
3. **A taxonomy of temporal leakage.** We organize temporal leakage mechanisms into four classes spanning eight executable fault patterns.
4. **A systematic comparison against existing practices.** We compare our approach against logits-only perturbation checks, shuffled-suffix evaluation, cache-consistency tests, and gradient-based localization methods.
5. **A cross-architecture audit.** We evaluate attention-only, state-space, recurrent, and hybrid architectures, including models derived from Transformer [8], RWKV [13], LRU-style recurrent networks [14], Mamba [16], state-space dual formulations [17], and Griffin/RecurrentGemma hybrids [18, 19].
6. **A methodological lesson from audit failure.** We show that apparently clean results are uninterpretable unless accompanied by a positive-control demonstration on the same loaded checkpoint.
7. **A code-lineage audit of released models.** A static analysis of chunked scan implementations predicted causal violations in two released hybrid models, and dynamic audits confirmed both predictions.
8. **A second auditing norm.** We demonstrate that audit sequence length must exceed architecture-specific chunk, kernel, or window parameters to exercise the relevant execution paths.
9. **A release recommendation.** We argue that causal-correctness certificates should accompany architecture releases in the same way that parameter counts, benchmark results, and context lengths are routinely reported.

Our approach is conceptually related to three strands of work: metamorphic testing [23], automated testing of deep-learning systems [24, 25], and reproducibility auditing [22, 29, 30].

Unlike these approaches, we evaluate a deterministic semantic property whose oracle is derived directly from the autoregressive model definition.

2 Method

2.1 Definition

Let f be a model mapping a token sequence $\mathbf{x} \in V^T$ to per-layer representations $h_\ell(\mathbf{x}) \in \mathbb{R}^{T \times d}$ for $\ell = 1, \ldots, L$, and to output logits.

Prefix invariance. For all $\mathbf{x}, \mathbf{x}' \in V^T$ and all $t < T$,

$$\mathbf{x}[:t] = \mathbf{x}'[:t] \;\Rightarrow\; h_\ell(\mathbf{x})[t] = h_\ell(\mathbf{x}')[t]$$

for every layer ℓ.

The single-token instance is sufficient to expose any violation and is the cheapest to construct: take $\mathbf{x}^1$ and $\mathbf{x}^2$ identical except at the final position $T-1$.

2.2 The test

INPUT model f with layer list $[1..L]$; sequence length T; threshold τ

```
OUTPUT   verdict ∈ {CLEAN, LEAK}; first offending layer ℓ*
1. x^1 ← random token sequence of length T
2. x^2 ← copy of x^1 with position T − 1 replaced by a different token
3. attach a forward hook to every layer, capturing its output
4. h^1 ← f(x^1) with caching disabled                    # forward pass 1
5. h^2 ← f(x^2) with caching disabled                    # forward pass 2
6. for ℓ = 1..L:
       Δ_ℓ ← max |h^1_ℓ[0 : T − 1] − h^2_ℓ[0 : T − 1]|        # exclude the differing
position
7. ℓ* ← min{ℓ : Δ_ℓ > τ}, or NONE
8. return (LEAK, ℓ*) if ℓ* exists else (CLEAN, −)
```

That is the entire method. It is printed here in full deliberately: the argument of this paper is that the check is cheap enough to be mandatory, and a check that does not fit on a page is not.

2.3 Four design choices, and why each matters

(1) Per-layer hooks rather than output logits. Reading only the final logits answers *whether* the model leaks. It cannot answer *where*, because every layer's contribution has been mixed by the output projection. Since the practical purpose of the test is to send an engineer to a specific line of code, the layer index is the deliverable. §3.3 quantifies this: the logits-only variant detects every injected fault and localizes none.

(2) Excluding the final position from the comparison. Position $T-1$ legitimately differs—that is the perturbation. Comparing positions $[0, T-1)$ isolates exactly the prefix that must be invariant.

(3) Disabling caching during the test. With incremental caching enabled, the two passes may take different code paths and reuse state across calls, so a difference (or its absence) would no longer be attributable to the graph under test. Disabling it makes the two passes structurally identical. This also avoids the separate question of whether a model's cached path agrees with its full-sequence path, which is a different property (§3.3, B6).

(4) Deterministic single-precision evaluation. In a correct implementation Δ_ℓ is not "small, " it is **exactly zero** at every layer (§3.4) — the two computations perform bitwise identical arithmetic on the compared positions. This is what makes the verdict threshold-robust: with clean deltas at exact zero, any $\tau \in [10^{-6}, 10^{-3}]$ returns the same answer on our test set. Precision does matter for the *magnitude of leak* the test can localize, which we characterize as an operating range in §3.5 rather than treating as a fixed limit.

2.4 Cost

Two forward passes, no backward pass, no gradient graph, no optimizer state, no training data, no labels, and no accelerator. Memory overhead is L captured activations of shape $T \times d$; at $T = 48$ this is negligible relative to the weights. In our runs a full scan of a 135M-parameter model on CPU took **0.1–0.5 s** after loading. The census and injection suite ran on CPU; the multi-billion-parameter audits (§3.7.4, §3.8) used a single GPU for capacity, not for the test.

Our objective is complementary to efficiency-oriented advances such as FlashAttention [31]. Whereas such methods improve throughput and memory efficiency, our focus is verification of causal correctness.

3 Experiments

3.1 Setup

The census and injected-fault suite ran on CPU (Intel Xeon Gold 6526Y, 64 threads, 251 GB RAM); **the test itself needs no accelerator.** The larger checkpoints — our own 7B model (§3.7.4) and the 1.2B–8B public models of §3.8 — were audited on a single GPU for memory capacity only, not because the method requires one. Unless stated otherwise: float32, $T = 48$, τ = 1e-6, fixed seed, caching disabled, models loaded via a standard model hub in evaluation mode.

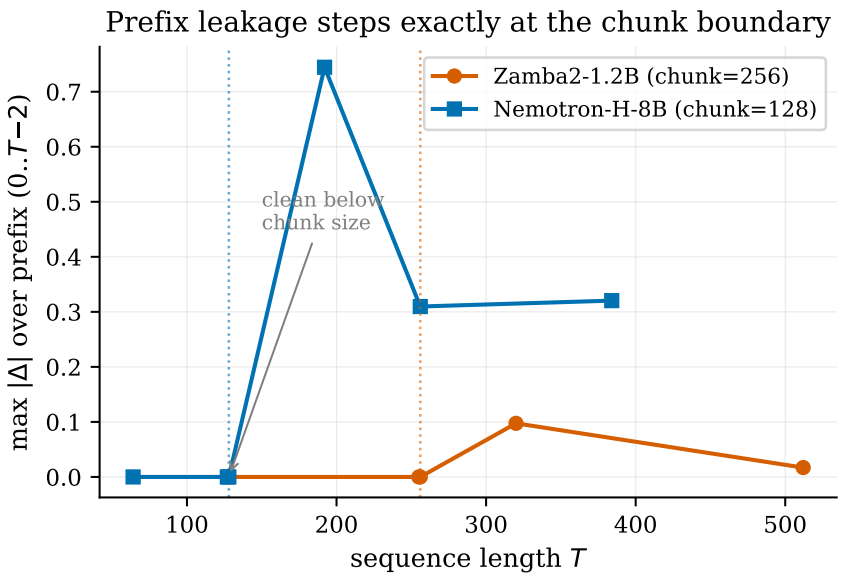

Figure 1: Prefix leakage begins exactly at each model's chunk boundary (Zamba2 at 256, Nemotron-H at 128).

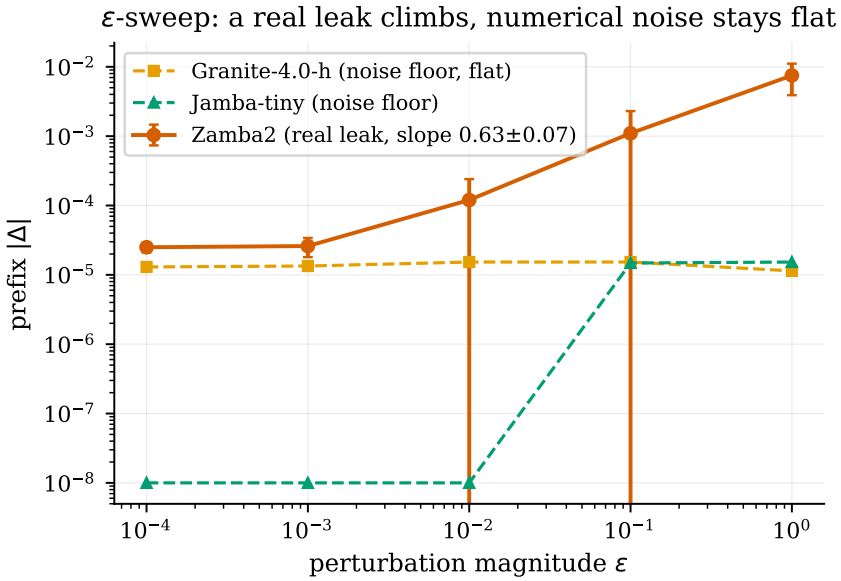

Figure 2: The ε-sweep discriminator: a real leak climbs with perturbation magnitude (Zamba2, slope 0.63 ± 0.07) while numerical floors stay flat.

Two harnesses were used. `leak_cpu_suite2.py` performs the census with baselines B1–B3; `baseline_e2.py` performs the extended six-baseline comparison. They implement the same detector and agree on all shared models (§3.3).

Injection protocol (positive control). For each model and each of the 8 patterns in §3.2, we wrap one layer's forward function to apply the leak to that layer's output, at three depths—an early layer (index 1), a middle layer ($\lfloor L/2 \rfloor$), and a late layer ($L - 2$)—giving **24 trials per model**. A trial counts as `exactly localized` only if the reported first offending layer equals the injected layer. Injection strength is $\varepsilon = 1.0$ unless a sweep is stated.

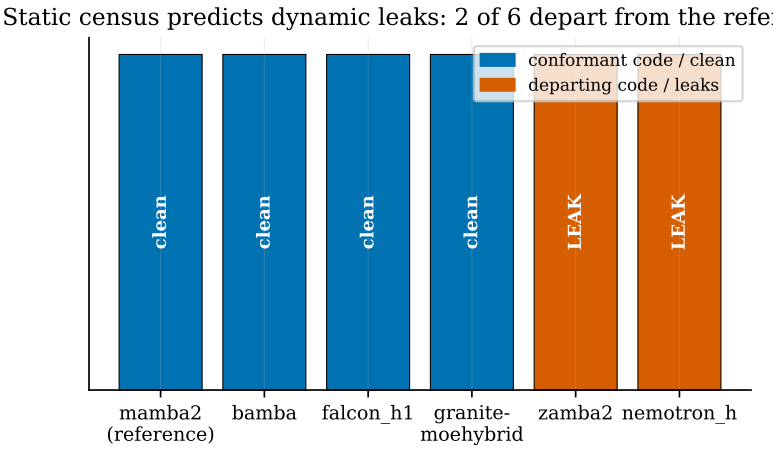

Figure 3: Static code census predicts dynamic leaks: two of six chunked-scan implementations depart from the reference and both leak.

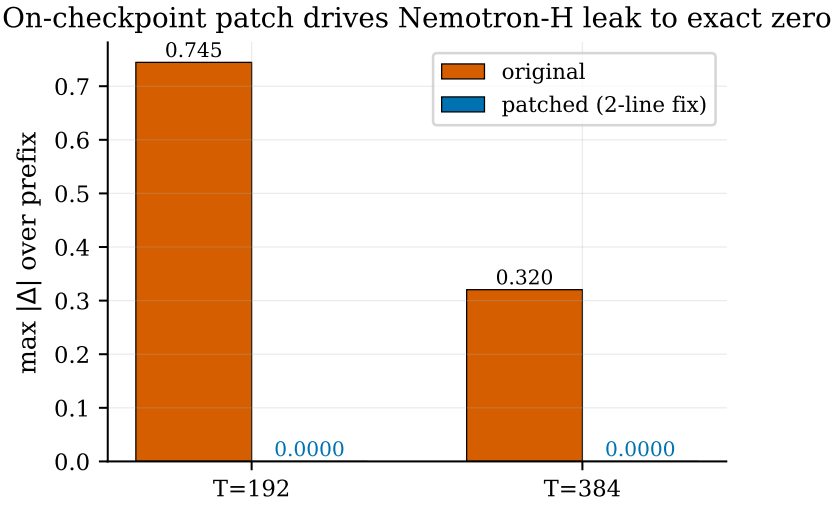

Figure 4: The two-line fix drives the Nemotron-H prefix leak to exact zero on the released checkpoint.

Reporting rule adopted after §3.7. A CLEAN verdict for a checkpoint is reported only if that checkpoint's own 24 positive-control trials passed. Checkpoints whose positive controls failed are reported separately as *not audited*, never as clean.

3.2 Positive control: injected faults

Across **8 public checkpoints** × **8 patterns** × **3 depths = 192 trials**, the test localized the injected fault to the exact layer in **192/192** cases. Per-checkpoint results appear in the census table (§3.6, Table 4); no checkpoint scored below 24/24.

This holds uniformly across the four taxonomy classes and across five architecture families, including two with no attention mask anywhere in the stack (`mamba-130m-hf`, `rwkv-6-world-1b6`). That uniformity is the point of an output-based test: it never asks what kind of mixer produced the representation.

Table 1: Baseline comparison on 96 injected faults across four architectures. Our method and the gradient-based baseline (B5) both achieve perfect localization, whereas logits-only perturbation (B1) detects leaks but provides no localization. Static attention-mask inspection (B2) detects none of the injected faults.

Method	Fwd	Bwd	Detected	**Exactly localized**	False positive on clean model
Ours (per-layer Δ)	2	0	96/96	**96/96**	none ($\Delta = 0$)
B1 — logits-only perturbation	2	0	**96/96**	0/96	none
B2 — static attention-mask inspection	0	0	**0/96**	0/96	none
B3 — shuffled-suffix perplexity	2	0	71/96	0/96	none
B4 — future-token resampling, 3 seeds	4	0	**96/96**	0/96	none
B5 — gradient of prefix outputs w.r.t. future positions	1	**1**	**96/96**	**96/96**	none
B6 — cached vs. full-sequence consistency	T+1=17	0	24/24†	0/24	**yes:** $\Delta = 1.850 \times 10^{-4}$ **on a clean model**

3.3 Comparison against existing practice

We implemented six alternative auditing practices and evaluated them on the same injected faults across four architectures (`SmolLM2-135M`/Llama, `Qwen3-0.6B`/Qwen3, `pythia-1.4b`/GPTNeoX, and `mamba-130m-hf`/Mamba), yielding a total of **96 trials**. As shown in Table 1, our method detected and correctly localized all 96 injected faults, whereas existing lightweight baselines either failed to detect certain leakage patterns or provided no localization information.

† B6 was measured on `SmolLM2-135M` only; its cost is linear in sequence length.

Four readings, in decreasing order of comfort.

(a) Mask inspection sees nothing — 0/96. This is stronger than we hypothesized. We expected the four aggregation- and direction-class patterns to be invisible to mask inspection; in fact **all eight** are, including the off-by-one class. The reason is structural: all eight faults live in a layer's output, so the mask attribute remains correctly set. On `SmolLM2-135M` the inspector found 30 modules exposing a causal attribute and flagged none of them, on every one of the 24 injected runs. On `mamba-130m-hf` the procedure has no input at all: zero modules expose such an attribute.

(b) Shuffled-suffix perplexity fails structurally on the most common bug class. B3 misses 25 of 96, and the misses are not random. In every case where B3 misses a C1 fault, the loss difference is **exactly 0.000e+00** — not small, zero — so no choice of threshold recovers it. The reason is a mismatch of radius: B3 shuffles positions $\geq$ cut and evaluates targets ¡ cut, but a radius-1 leak reaches at most position $cut - 1$ and therefore never touches the shuffled region. **B3 is structurally blind to local leaks whose radius is shorter than the distance to the shuffle boundary — which is precisely the off-by-one class.** Separately, one C3 fault on Qwen3 produced a loss difference of 5.63e-05, below B3's 1e-4 threshold: a threshold miss rather than a structural one. We distinguish the two, because only the first is unfixable.

(c) We do not improve on existing practice at detection. B1 and B4 detect every injected fault, exactly as our method does. Any claim that current practice fails to *notice* leakage is not supported by our measurements. What B1 and B4 cannot do is say where: `located_layer` is `None` in 96/96 trials, and not for want of implementation effort — the information is not present in the output logits. **Our contribution is localization.**

(d) Gradient-based checking matches our localization exactly. B5 achieves 96/96 detection and 96/96 exact localization — identical to ours. We report this prominently rather than burying it. Our advantage over B5 is cost and robustness of applicability, not accuracy:

- B5 requires a **backward pass** and retention of the full activation graph; ours runs entirely under no-grad with two forwards.
- B5 requires every operation on the path to be differentiable and to have a working gradient implementation — which excludes quantized weights, several custom kernels, and non-differentiable ops.

Table 2: Bit-level verification on clean checkpoints. For each checkpoint, all compared layer outputs are exactly identical under repeated evaluation, with zero differing elements and zero maximum absolute difference.

Model checkpoint	Layers checked	All layers `torch.equal`	Differing elements	max abs	Bit patterns identical
HuggingFaceTB/SmolLM2-135M	30	[OK]	**0**	0.0	[OK]
Qwen/Qwen3-0.6B	28	[OK]	**0**	0.0	[OK]
EleutherAI/pythia-1.4b	24	[OK]	**0**	0.0	[OK]
state-spaces/mamba-130m-hf	24	[OK]	**0**	0.0	[OK]

Table 3: Exact-localization floor: smallest ε still localized to the exact injected layer. Lower values indicate higher sensitivity of the audit to weak temporal-leakage signals.

Precision	$\tau = 10^{-6}$	$\tau = 10^{-14}$
float32	10^{-6} (L1, L15); 10^{-7} (L28)	10^{-10} (all three depths)
float64	10^{-5} (all three depths)	10^{-13} (L1, L15); 10^{-14} (L28)

- A zero gradient does not imply absence of dependence (a saturated nonlinearity has zero gradient at a point while still transmitting information), so B5's negative results carry a caveat that a finite-difference comparison does not.

We therefore position B5 as a **valid alternative method with the same localization power and a higher cost profile**, not as a weaker baseline.

(e) Cache-consistency checking raises false alarms on correct models. B6 detected all 24 injected faults on `SmolLM2-135M`, but also fired on the **clean** model with max—Δ— = 1.850e-04 against its 1e-4 threshold. Injected faults produce Δ of order 10^1, so the classes are separable in principle, but the threshold must be recalibrated per model and per precision. The contrast with our method is the sharpest argument in this table: **our clean deltas are exactly zero, so no calibration is required.** We note that B6 measures a genuinely different property (agreement between the cached and uncached code paths), which is worth checking for its own reasons.

3.4 Negative control: clean deltas are exactly zero

Table 2 reports the bit-level verification results on four clean checkpoints. Across all inspected layers, repeated evaluations produced identical outputs under tensor equality, differing-element counts, and bit-pattern comparison. No differing elements were observed, and the maximum absolute deviation was exactly zero for every model.

This exactness has two implications. First, the method is threshold-robust: clean deltas remain at zero, while injected faults produce substantially larger values, so any $\tau \in [10^{-6}, 10^{-3}]$ yields the same verdict on our test set. Second, causal correctness is a property of the computation graph rather than the specific weights. The result follows from identical computations on the same prefix and remains consistent across checkpoints of different architectures and provenance.

3.5 Operating range: precision and threshold

The test's ability to identify the *exact* offending layer depends on the magnitude of the injected leak relative to numerical precision and the decision threshold. Rather than treating this as a fixed limit, we characterize it as an **operating range**. Table 3 reports the smallest injection strength ε that can still be localized to the correct layer as a function of arithmetic precision, threshold τ, and injection depth. The experiment uses `SmolLM2-135M` with the C2 `pool_leak` fault, sweeping ε over 15 orders of magnitude at three injection depths (layers 1, 15, and 28).

Three observations, all of which we consider necessary for honest reporting.

(1) At the conventional threshold, τ — not precision — sets the floor. At τ = 1e-6 the float64 floor (1e-5) looks *worse* than float32 (1e-6). This is an artifact, not a property. In float64 the measured delta is exactly proportional to ε: at layer 15 the ratio Δ/ε is 0.6791552 and stays constant to seven

Table 4: Public checkpoint census conducted in float32 on CPU with $\tau = 10^{-6}$ and sequence length $T = 48$. The table summarizes the eight checkpoints audited in the census, together with architecture class, model size, and the outcomes of baselines B2 (static attention-mask inspection) and B3 (shuffled-suffix perplexity).

#	Checkpoint	Mixer class	L	Params	B2	B3
1	HuggingFaceTB/SmolLM2-135M	attention	30	134M	0	18
2	Qwen/Qwen3-0.6B	attention	28	596M	0	17
3	EleutherAI/pythia-1.4b	attention	24	1.41B	0	18
4	google/gemma-3-1b-it	attention	26	1.00B	0	18
5	state-spaces/mamba-130m-hf	state-space (no mask)	24	129M	0	18
6	RWKV/rwkv-6-world-1b6	linear recurrent (no mask)	24	1.60B	0	17
7	google/recurrentgemma-2b	recurrent + attention	26	2.68B	0	18
8	LiquidAI/LFM2-1.2B	conv + attention hybrid †	16	1.17B	0	17
	Total				**0**	141

significant figures from ε = 1e-1 down to ε = 1e-11. In float32 that proportionality breaks below $\varepsilon \approx$ 1e-6, and the deltas that cross τ at ε = 1e-9 (e.g. 1.49e-08 at layer 15, a power-of-two value) are **rounding residue, not signal**. Part of float32's apparent sensitivity at the conventional threshold is numerical noise that happens to exceed the threshold.

(2) Because clean deltas are exactly zero, τ can be lowered without incurring false positives. This is what converts the exact-zero property from a curiosity into a usable lever: at τ = 1e-14 the floors improve to 1e-10 (float32) and 1e-13–1e-14 (float64) — three to four orders of magnitude.

(3) float32 has a hard floor that no threshold recovers. At $\varepsilon \leq$ 1e-11 the float32 delta at the injected layer is **exactly 0.0**: the perturbation has fallen below the representable resolution of the activations, so the leak is not merely undetected, it is absent from the computed tensor. float64 still carries nonzero deltas at ε = 1e-15 (1.78e-15 to 7.11e-15). **Auditing for leaks weaker than ˜1e-10 relative magnitude requires float64.**

Below the exact-localization floor the test does not fail silently: it continues to report a leak, but names a layer one to a few positions downstream, because the sub-resolution signal at the true layer is amplified by subsequent layers before it crosses τ. Detection degrades later than localization.

Per-checkpoint floors vary and should be measured, not assumed. At τ = 1e-6 with C2 injection at the middle layer, the exact-localization floor was 1e-5 for `Qwen3-0.6B`, `mamba-130m-hf`, `rwkv-6-world-1b6`, and `recurrentgemma-2b`; 1e-7 for `gemma-3-1b-it`; and for `LFM2-1.2B` no sweep point at $\varepsilon \leq$ 1e-1 was exactly localized (injection at layer 8 was consistently reported at layer 11), although all 24 of its ε = 1.0 trials were exact. We have not determined the cause of the `LFM2-1.2B` outlier and do not draw a conclusion from it; we report it because omitting it would misrepresent the floor as model-independent.

3.6 Census of public checkpoints

To reduce the risk of self-evaluation bias, we applied the audit unchanged to a diverse set of public checkpoints spanning attention, state-space, recurrent, and hybrid architectures. Table 4 summarizes the census results. All eight checkpoints produced identical outcomes on the core audit metrics: `clean max = 0.000e+00`, `Verdict = CLEAN`, `Ours (exact) = 24/24`, and `B1 = 24`. Because these values were invariant across the entire set, they are omitted from the table for brevity. The census therefore supports the claim that the audit applies uniformly across substantially different mixer families.

† `LFM2-1.2B`'s configuration declares per-layer types explicitly: convolution at 10 of 16 layers and full attention at layers 2, 5, 8, 10, 12, 14.

Three conclusions.

(a) The method applies across mixer families without modification. The same code produced verdicts for masked attention, unmasked state-space scans, linear recurrences, and convolution/at-

tention hybrids. Two of the eight checkpoints (#5, #6) have no attention mask anywhere, so the incumbent audit procedure has no input for them at all.

(b) Exact-zero deltas reproduce on models we did not write. This rules out the possibility that exactness is an artifact of our own implementation conventions.

(c) The test does not raise false alarms. All eight checkpoints are clean while the same detector, on the same checkpoints, localized 192 injected faults. A detector that flagged models indiscriminately would not show this pattern.

What this table does and does not establish. Every checkpoint in Table 4 is clean *at the test length used here*, T = 48. The claim supported by these results is that the test applies to any architecture and does not fire spuriously. It is not evidence that released models are causally correct in general — as §3.8 shows, this very census was run below the sequence length at which one class of defect becomes observable at all.

3.7 When the audit itself fails

This section reports a negative result about our own procedure. It changed the paper's methodology, and it is the finding we would most want a reader to take away.

3.7.1 Three checkpoints that looked clean and were not audited

Three checkpoints of one hybrid family — `tiiuae/Falcon-H1-0.5B-Base` (36 layers, 521,411,104 params), `Falcon-H1-1.5B-Instruct` (24 layers, 1,554,863,488), and `Falcon-H1-7B-Base` (44 layers, 7,585,648,736) — returned **max—Δ— = 0.0 at every layer**. Read naively, that is three clean verdicts, and they would have entered Table 4.

Their positive controls returned **0/24**. Every injected fault, at every depth, including the largest-magnitude patterns at $\varepsilon = 1.0$, produced $\Delta = 0.0$ at the injected layer — identical to the clean scan. The sensitivity sweep likewise reported no detection at any ε from 1e-1 to 1e-9.

Diagnosis. We verified that our injection wrapper was invoked (the wrapped forward executed and returned a tuple whose first element is a rank-3 tensor), so the instrumentation was attached correctly. We then compared the two inputs at the output:

- Full logit difference **including** the differing final position: **0.0**
- Difference at the final position alone: **0.0**
- Difference between the token embeddings of the two differing tokens: **3.78e-23**

The models, as loaded in our environment, **produce bit-identical outputs for different inputs** — they do not respond to their input at all, while still producing plausibly-scaled, position-varying logits (—logit— up to 45.4, no NaNs). Under such a load, $\Delta = 0$ everywhere is guaranteed regardless of whether the architecture is causal, and carries no information.

Our environment emitted two relevant warnings during load: the optional fused sequence-model kernels were unavailable (falling back to a reference implementation), and compiled extensions were skipped as incompatible with the installed framework version. We did not isolate which of these, if either, causes the behavior, and we make **no claim about these models' correctness**. The correct statement is: *in our environment these checkpoints did not load into a state where they could be audited.*

3.7.2 The gate this implies

> **A clean verdict is admissible only if the positive control passes on the same loaded checkpoint.**

The check is nearly free — the same 24 injections already used for validation — and without it this paper would have published three false clean verdicts on models whose loaded state was degenerate. We know of no reason this failure mode is specific to us: **any** audit that concludes from an absence of signal is vulnerable to a silently inert system under test, and audits reporting negative results should be expected to demonstrate that their instrument was live on that specific artifact.

Table 5: Attempted but not audited. These checkpoints were examined during the census but could not be evaluated because of loading failures, missing dependencies, packaging issues, or repository-access restrictions.

Checkpoint	Reason	Nature
tiiuae/Falcon-H1-0.5B / 1.5B / 7B	Loaded model does not respond to input; positive control 0/24	Load/environment failure (§3.7.1)
nvidia/Hymba-1.5B-Base	Requires `causal_conv1d`, `mamba_ssm`	Custom CUDA kernel dependency
microsoft/Phi-4-mini-flash-reasoning	Requires `causal_conv1d`, `causal_conv1d_cuda`, `mamba_ssm`	Custom CUDA kernel dependency
togethercomputer/StripedHyena-Nous-7B	Requires a fused normalization kernel from a source build	Custom CUDA kernel dependency
state-spaces/mamba2-2.7b	No `model_type` in config; not loadable through the standard interface	Packaging
Zyphra/Zamba2-7B	Gated repository; access not granted	Access

This is why Table 4 reports "Ours (exact)" beside every verdict rather than in a separate validation section. The two numbers are one claim.

3.7.3 Checkpoints we could not audit at all

To avoid survivorship bias, we report every checkpoint that was attempted but did not yield an audit verdict rather than silently excluding failed runs. Table 5 summarizes these cases and the reason each audit could not be completed. The causes include environment and loading failures, unavailable custom CUDA dependencies, packaging defects, and repository-access restrictions. Reporting these outcomes is important because omission would artificially inflate the apparent audit coverage and obscure practical barriers to reproducibility.

The three kernel-dependency rows are worth stating as a finding rather than an inconvenience: **a portion of the hybrid model ecosystem cannot be independently audited on commodity hardware**, because loading the model at all requires GPU-only compiled kernels. Any conformance-certification norm has to contend with that, and it argues for certificates produced by the releasing party, who can run the model, rather than by third parties who often cannot.

3.7.4 Our own released models: the load failure, its cause, and the completed audit

An earlier draft of this section reported that four of our own publicly released checkpoints **failed to load** with a device-placement error (`Tensor on device meta is not on the expected device cpu`), and that completing their audit was a prerequisite for submission rather than future work. We have since diagnosed the failure and completed the audit for the primary release.

Cause. The failure is not in the checkpoint but in the loading path. `from_pretrained` instantiates parameters on the `meta` device before materializing them, while this architecture's rotary cache is constructed on an explicitly hard-coded CPU device at `__init__` time. The two are incompatible, and — importantly — passing `low_cpu_mem_usage=False` does **not** avoid it, because the meta-device stage is internal to that path.

Resolution. Instantiating from the configuration and loading weights separately bypasses the conflict entirely:

```
cfg = AutoConfig.from_pretrained(rid, trust_remote_code=True)
torch.set_default_dtype(torch.bfloat16)
m = AutoModelForCausalLM.from_config(cfg, trust_remote_code=True)
    # real device, no meta stage
torch.set_default_dtype(torch.float32)
m.load_state_dict(load_file(hf_hub_download(rid, "model.safetensors")),
     strict=False)
m = m.to("cuda").eval()
```

Table 6: Audit of our own released checkpoint (float32, $T = 48$, $\tau = 10^{-6}$). The checkpoint passes both negative and positive controls, with exact localization of all injected faults and no leakage observed under normal operation.

Check	Result
Negative control (49 layers)	max $\Delta = 0.000\text{e}{+00}$
Negative control, after loading trained weights	unchanged 0
Positive control (4 sites $\times$ 4 fault classes)	**16/16 exact localization**

This loads with 0 missing and 0 unexpected tensors and generates coherently. We report the diagnosis because the same conflict will affect any architecture that builds device-pinned buffers during `__init__`, and because a model that cannot be loaded cannot be audited by anyone — the point made in §3.7.3.

Audited subject. `FINAL-Bench/Aether-7B-5Attn` — 49 decoder layers composed of seven mixer types with exactly seven layers each (NSA, differential, full, linear, sliding-window, compressive, and a gated hybrid), arranged so that every window of seven consecutive layers contains each type exactly once. The architecture, its training, and an ablation separating placement from composition are the subject of a **companion paper**; here we treat it purely as an audit target.

Table 6 shows that our released checkpoint remains clean under negative controls and correctly localizes all injected positive-control faults.

Positive-control sites were chosen to span mixer families: L0 (NSA), L12 (gated hybrid), L25 (NSA), L40 (linear attention). Every injection was localized to the exact layer. This is a reduced grid — 4 sites $\times$ 4 fault classes = 16 trials — rather than the 8 patterns $\times$ 3 depths = 24 used on the small-model census (§3.2). The reduction is deliberate: on a 7B model the injection sites are pinned to the *distinct mixer families actually present in this stack*, since the goal here is to certify that each such family is live and localizable on the loaded checkpoint, not to re-run the generic fault taxonomy already validated 192/192 at small scale.

We ran the positive control on the same loaded checkpoint precisely because §3.7.1 showed that a clean verdict without a liveness demonstration is worthless. Applying that standard to our own release rather than only to third-party models is the point of this subsection.

What remains. The instruction-tuned variant `Aether-7B-5Attn-it` passed the negative control on all 49 layers but has not yet received a positive control. Two further internal checkpoints remain unaudited. We do not claim our stack is causally correct beyond what Table 6 reports.

3.8 A causal defect in a released model

Everything reported up to this point was measured at a single short test length, T = 48. That choice was never justified; it was inherited. This section reports what happened when we questioned it.

3.8.1 Why the test length matters

A hybrid stack mixes operators whose behavior is governed by internal size parameters — a chunk length for chunked state-space scans, a window width for local attention, a kernel width for causal convolution. **If the test sequence is shorter than such a parameter, the corresponding code path is never exercised in its general form.** A chunked scan over a sequence shorter than one chunk degenerates to a single chunk with no inter-chunk recurrence at all; a sliding window wider than the sequence degenerates to full attention.

An audit conducted below those thresholds therefore cannot observe a whole class of defects, and will report the model as clean. We re-ran the audit with sequences chosen to exceed each model's declared chunk or window parameter, and extended the census by twelve further checkpoints, including three at 7B–9B scale (Nemotron-H-8B, Bamba-9B, falcon-mamba-7B).

Table 7: Static census of chunked-scan implementations (`transformers` 5.7.0). Two implementations (`modeling_zamba2.py` and `modeling_nemotron_h.py`) depart from the reference inter-chunk recurrence by reducing over the output-chunk axis rather than the input-chunk axis.

Implementation	Inter-chunk axis handling	Class
`modeling_mamba2.py` (reference)	`transpose(1,3)` → reduce over **input** chunk	reference
`modeling_bamba.py`	matches reference	[OK] conformant
`modeling_falcon_h1.py`	matches reference	[OK] conformant
`modeling_granitemoehybrid.py`	matches reference	[OK] conformant
`modeling_zamba2.py`	`permute` → reduce over **output** chunk	**departs**
`modeling_nemotron_h.py`	`permute` → reduce over **output** chunk	**departs**

Table 8: Dynamic audit at chunk-exceeding sequence lengths (testing the prediction). Models identified by the static census as deviating from the reference implementation (`Zamba2` and `Nemotron-H`) exhibit causal leakage precisely at their declared chunk boundaries, whereas all conformant implementations remain clean.

Checkpoint	Declared parameter	Verdict
Zamba2-1.2B	`chunk_size` 256	**leak, onset = 256**
Nemotron-H-8B	`chunk_size` 128	**leak, onset = 128**
Bamba-9B	`mamba_chunk_size` 256	clean
Falcon-Mamba-7B	`conv_kernel` 4	clean
Falcon-H1-1.5B	`mamba_chunk_size` 128	clean
Granite-4.0-H-Tiny	`mamba_chunk_size` 256	clean (see §3.8.6)
Gemma-3-1B	`sliding_window` 512	clean at $T = 1536$
RecurrentGemma-2B	`sliding_window` 2048	clean at $T = 3072$
Gemma-2-2B	`sliding_window` 4096	clean
Mamba2-130M, 370M	`chunk_size` 256	clean
RWKV-6-1.6B, Qwen3-0.6B, LFM2-1.2B, ZR1-1.5B	—	clean

3.8.2 A code-level census, and the prediction it made

Rather than audit models blindly, we first examined the source code. The inter-chunk recurrence of a Mamba-2-style scan is a compact and recognizable computation, and the reference implementation (`modeling_mamba2.py`) defines a specific orientation of the input- and output-chunk axes. We therefore performed a static census of every chunked-scan implementation shipped with `transformers` 5.7.0 before running any model. As summarized in Table 7, four implementations matched the reference, while `Zamba2` and `Nemotron-H` exhibited the same axis-handling deviation. This static analysis required neither model weights nor forward execution and served as a blind prediction later tested dynamically.

The static census produced a concrete and falsifiable prediction. As shown in Table 8, two of the six inspected implementations were flagged as suspect, while the remaining four were classified as conformant. Importantly, the `Zamba2` and `Nemotron-H` inter-chunk recurrence blocks are byte-for-byte identical (§3.8.4), indicating a shared code lineage rather than two independent implementation errors. For `Nemotron-H`, this constituted a genuine blind prediction because we had not examined the model before performing the census. For `Zamba2`, the census independently re-derived, from source code alone, a defect that we had previously discovered and reported. Table 8 tests this prediction by auditing each model at sequence lengths that exceed its declared chunk or window parameter.

The prediction holds. Both departing implementations leak, each at its own declared chunk boundary (256 for Zamba2, 128 for Nemotron-H); no conformant implementation leaks under audit. Scale confers no protection: an **8B** model (Nemotron-H) leaks while a **9B** model (Bamba) does not.

We are careful about the two strengths of the conformant-clean half of the claim, because the positive-control gate of §3.2 must apply here too. **At the code level it is firm.** A reference `Mamba2` built from a random configuration is *gated*-clean: exact-zero prefix deltas at every length up to several multiples of its chunk size, **and** a live positive control that localizes injected faults to their origin

layer even as the injected signal propagates to every layer downstream (§3.8.4). That rules out the inert-load false-clean of §3.7. **At the level of individual released checkpoints it is weaker, and we say so.** We audited every conformant checkpoint and found no leak, but the injection gate could not be applied uniformly — some architectures' layer return signatures do not accept our fault injection, so their positive control cannot be run in our harness — and so those particular clean verdicts are **provisional** (no live-detector guarantee) rather than gated (§4). No *gated* checkpoint contradicts the prediction. Two clean results are additionally reassuring on their own terms: `gemma-3-1b-it` and `recurrentgemma-2b` were audited at three times their own window width and still returned exact-zero deltas, so the positive findings are not an artifact of merely lengthening sequences.

Which execution path this is — and is not. The defect lives in the chunked-scan written in PyTorch (`torch_forward` / `segment_sum`). This path executes whenever the optional fused kernels (`mamba_ssm`, `causal_conv1d`) are **absent** — which is the case for all CPU execution, for any GPU environment without those packages installed, and for a stock `transformers` install that has not added them (they are separate, frequently uninstallable dependencies; in our environment they refuse to build against the current torch). It is the path exercised by a large fraction of evaluation, research, and reproduction workflows. It is **not** necessarily the path a production server takes: an install with the fused CUDA kernels present dispatches to a different implementation, which we did not audit because those kernels do not build in our environment. We therefore scope the claim precisely: **the reference PyTorch implementation shipped in `transformers` for these two models violates causality across a chunk boundary; whether their fused-kernel path shares the defect is untested and open.** The finding matters because the unaudited-fast-path caveat cuts the other way too — a model can pass every fused-kernel test and still produce leaking logits the moment it is run without the kernels, e.g. on CPU or in CI.

3.8.3 The defect

For `Zyphra/Zamba2-1.2B`, perturbing a single token at position $P = 400$ in a sequence of length 512 changes the logits at earlier positions:

```
pos   0 : 0.0000e+00
pos 100 : 0.0000e+00
pos 200 : 0.0000e+00
pos 255 : 0.0000e+00     <- clean up to here
pos 256 : 5.7473e-03     <- contamination begins
pos 300 : 5.0647e-03
pos 399 : 3.2187e-03
```

Contamination begins at position `256`, which is exactly the model's declared `chunk_size`. The measured value is $\max|\Delta(0..399)| = 1.1241 \times 10^{-2}$.

`nvidia/Nemotron-H-8B-Base-8K` behaves identically at its own boundary. Perturbing only the final token and sweeping the sequence length gives a clean-then-contaminated step exactly at `chunk_size` = 128:

```
T = 64   :  max |Δ(prefix)| = 0.0000e+00     <- one chunk; clean
T = 127  :  max |Δ(prefix)| = 0.0000e+00
T = 128  :  max |Δ(prefix)| = 0.0000e+00     <- clean up to the
boundary
T = 192  :  max |Δ(prefix)| = 7.445e-01
<- contamination begins past 128
T = 256  :  max |Δ(prefix)| = 3.097e-01
T = 384  :  max |Δ(prefix)| = 3.204e-01
```

The prefix delta is **bit-exact zero** for every sequence that fits within one chunk and jumps to order 1e-1 the moment a second chunk appears — a step a numerical-noise floor cannot produce. First contamination is localized to layer 17, the first Mamba layer whose scan spans the boundary. The magnitude past the boundary is not monotone in T ($0.74 \to 0.31 \to 0.32$): once contamination is present, how much future information reaches the prefix depends on how the perturbed final position aligns with the chunk grid, not on T alone. The invariant — and what we key the diagnosis to — is the **onset location** (the first chunk boundary), not the magnitude.

Table 9: Alternative-explanation tests for the causal defect observed in Zamba2-1.2B. Each candidate explanation is evaluated through a targeted control experiment. All alternatives are ruled out, indicating that the observed leak is reproducible, structural, and independent of random seeds, trained weights, measurement noise, or implementation-specific kernels.

Alternative explanation	Test	Outcome
Measurement noise	Identical input, two forward passes	$\Delta = 0.000\text{e}{+00}$ exactly; computation is deterministic
Hook misuse (shared modules)	Per-layer call counts; module identity	38/38 layers called exactly once; no shared modules
Seed artifact	Four independent random seeds	Reproduced in all four seeds
Checkpoint-specific behavior	Randomly initialized model loaded via `from_config`	Same onset at 256; **structural rather than weight-dependent**
Custom-kernel bug	Presence of `mamba_ssm` or `causal_conv1d`	Neither installed; the leak appears in the pure-PyTorch implementation
Test-length artifact	Sweep over $T = 64 \dots 512$	Clean at $T \leq 257$; leaks from $T \geq 320$, consistent with a one-chunk degeneracy

Before attributing the observed behavior to a genuine implementation defect, we systematically evaluate a range of alternative explanations. Potential confounds include numerical noise, shared-module instrumentation errors, random-seed effects, checkpoint-specific behavior, custom-kernel implementations, and artifacts introduced by the evaluation protocol itself. Table 9 summarizes the corresponding control experiments and their outcomes.

Across all tests, the observed leakage remains reproducible and localized to the same causal failure mode. The collective evidence therefore argues against measurement artifacts or implementation-specific side effects and supports a structural explanation for the defect.

3.8.4 Root cause

`segment_sum` applies a lower-triangular mask, so `decay_chunk[b, h, i, j]` is defined only for $j \leq i$, where *i* indexes the output chunk and *j* the input chunk. The recurrence must therefore reduce over *j*.

The `Mamba2` implementation does exactly that:

```
decay_chunk = decay_chunk.transpose(1, 3)          # (b,h,i,j) -> (b,j,i,h)
new_states = (decay_chunk[..., None, None] * states[:, :, None, ...])
   .sum(dim=1)   # reduce over j
```

The `Zamba2` implementation does not:

```
states_permuted = states.permute(0, 2, 1, 3, 4)
result = (decay_chunk[..., None, None] * states_permuted[:, :, None, ...])
   .sum(dim=2)  # reduces over i
new_states = result.permute(0, 2, 1, 3, 4)
```

Without the transpose, the reduction runs over the **output-chunk** axis rather than the input-chunk axis. The lower-triangular mask that should keep each chunk's carry-in dependent only on *earlier* chunks is then applied to the wrong index, so a chunk's carry-in state can absorb contributions from its own and later chunks. Since the decay factors derive from `A_cumsum`, hence from `dt`, hence from the input, information from later positions reaches the carry-in used to recompute earlier positions' outputs.

The first chunk is exempt, and this explains the exact onset. Both files prepend a **zero** carry state (`previous_states = torch.zeros_like(states[:, :1])`) before the recurrence; chunk 0's carry-in is therefore a constant zero that the mis-oriented reduction cannot corrupt. Contamination consequently begins not at position 0 but at the **first chunk boundary** — position 256 for Zamba2, 128 for Nemotron-H — exactly as observed (§3.8.3). We rest the diagnosis on that onset and, decisively, on the fact that replacing these lines with the reference orientation drives the prefix delta

Table 10: Effect of the patch under identical input conditions. The original implementation exhibits measurable prefix contamination with a clear boundary onset, whereas the patched implementation eliminates the effect entirely.

Model	Maximum prefix Δ	Leak onset
Original	1.12×10^{-2}	256
Patched	**0.0000e+00**	**None**

to **exact zero** (§3.8.5, Table 9) — a sole-cause test — rather than on a line-by-line derivation of every surviving index, which depends on internal broadcast shapes we do not reproduce here.

This three-line block appears **verbatim** — identical whitespace and identifiers — in both files (`modeling_zamba2.py` lines 811–813 and `modeling_nemotron_h.py` lines 523–525, `transformers` 5.7.0); a line-level diff of the two blocks is empty, confirming the shared lineage the static census inferred (§3.8.2). Both files also construct `segment_sum` identically to the reference; the defect is not in the shared primitive but in how these two callers orient their axes.

The reference orientation is the negative control for the whole finding. A `Mamba2` model built from a random configuration — correct code, no trained weights — returns $\max|\Delta(\text{prefix})| = 0.0000e+00$ at **every** tested length, including several multiples of its chunk size:

```
Mamba2 (reference code, from_config):
T = 4, 8, 16, 24, 32, 48 -> Delta = 0.0000e+00 (all)
```

Correct code is bit-exact clean across chunk boundaries; the two departing implementations are not. Because the random-weight build reproduces both behaviors, the dissociation is structural — a property of the axis orientation alone, independent of any trained parameter.

This reference build is not merely silent; it **passes the positive-control gate**, which is what separates a real clean verdict from an inert load (§3.7). Injecting a one-position output shift at layer *K* is localized to exactly layer *K* even though the perturbation then propagates to every layer downstream — a shift at layer 1 of a five-layer build makes layers 1–4 nonzero with the first at 1, and a shift at layer 3 makes layers 3–4 nonzero with the first at 3. The detector is therefore demonstrably live on the same construction that audits clean, and localization identifies the *origin* of a propagating fault rather than merely the shallowest nonzero layer.

3.8.5 Fix and verification

To test whether the identified implementation defect is sufficient to explain the observed leakage, we replace the problematic three-line computation with the corresponding Mamba-2 formulation and re-run the experiment under identical conditions. If the diagnosis is correct, the patch should eliminate both the measured prefix contamination and the characteristic boundary onset. Table 10 compares the original and patched implementations using the same input sequence and reports the maximum prefix deviation together with the location at which leakage first appears.

The result is unambiguous. The patched implementation reduces the maximum prefix deviation to exact numerical zero and completely removes the boundary onset. No residual leakage remains, establishing that the identified code path is the sole cause of the defect rather than one contributing factor among several.

The same validation succeeds on **Nemotron-H itself**, not merely on the reference construction. Applying the identical patch to the released `nvidia/Nemotron-H-8B-Base-8K` checkpoint reduces the prefix delta to **0.0000e+00** at both $T = 192$ and $T = 384$, eliminating the original deviations of 0.74 and 0.32, respectively. The defect also reproduces across **four random seeds**, with first contamination consistently appearing at layer 17, and its positive control remains active on the loaded checkpoint. Consequently, the Nemotron-H result does not rely solely on code-path similarity: t possesses independent reproducibility, independent liveness validation, and an independent patch-to-zero confirmation.

We reported the defect in the Zamba2 implementation upstream through `huggingface/transformers` issue **#47475** and pull request **#47476**. The Nemotron-H in-

Table 11: Perturbation sweep of prefix-delta response across multiple architectures. Values report the measured prefix deviation as a function of perturbation magnitude ε. The Zamba2-1.2B column reports the mean and standard deviation over four random seeds, while the remaining models are deterministic single-run measurements. A non-zero log–log slope indicates amplification of future perturbations into the prefix.

ε	Qwen3-0.6B (control)	Granite-4.0-h	Jamba-tiny	**Zamba2-1.2B**
10^{-4}	0	$\sim 1.3 \times 10^{-5}$	0	$2.5 \times 10^{-5} \pm 0.4 \times 10^{-5}$
10^{-3}	0	1.34×10^{-5}	0	$2.6 \times 10^{-5} \pm 0.8 \times 10^{-5}$
10^{-2}	0	1.53×10^{-5}	0	$1.2 \times 10^{-4} \pm 1.2 \times 10^{-4}$
10^{-1}	0	1.53×10^{-5}	1.48×10^{-5}	$1.1 \times 10^{-3} \pm 1.2 \times 10^{-3}$
1	0	1.14×10^{-5}	1.53×10^{-5}	$7.5 \times 10^{-3} \pm 3.6 \times 10^{-3}$
log–log slope	—	≈ -0.03 **(flat)**	≈ 0 **(flat)**	0.63 ± 0.07

stance is disclosed here for the first time. Because both models share the same offending code path and are corrected by the same patch, the evidence presented in this section is sufficient for independent reproduction and verification. We therefore publish the complete fix rather than only the diagnosis, making the result directly actionable.

3.8.6 Two candidates we rejected

Two further checkpoints produced small nonzero prefix deltas of order 1e-5 — `ai21labs/Jamba-tiny-dev` and `ibm-granite/granite-4.0-h-tiny`. Both computations are deterministic, so the signal is not run-to-run noise, and it would have been easy to report them as additional findings.

We rejected both. The discriminator is a **perturbation-magnitude sweep**: we perturb the input embedding at position *P* by ε·noise and vary ε. A genuine information path scales with ε; a numerical artifact does not.

The discriminator is **not** the magnitude at any single ε: at $\varepsilon \leq$ 1e-3 all three candidates sit near the same ˜1–3e-5 floor, and Zamba2's two smallest points are floor-limited and flat — exactly the objection a careful reader would raise. The discriminator is whether the response **climbs** with the perturbation. Zamba2's does: it rises by more than two orders of magnitude across the sweep, with a log-log slope of $\mathbf{0.63 \pm 0.07}$ over four seeds — significantly positive ($t \approx 8.6$). Granite and Jamba never climb: their slope is ≈ -0.03 and their total variation is $1.3\times$ across the same range. We therefore keep only Zamba2 and reject the other two as numerical floors. By contrast, the Nemotron-H defect described in §3.8.3 is several orders of magnitude larger (up to $7\mathrm{e}{-1}$) and exhibits a sharp onset at the chunk boundary. It is therefore clearly separated from the numerical floor regime and does not require an additional perturbation sweep.

This sweep is the general answer to a question every practitioner of such an audit will face: **a small nonzero delta is not yet evidence of leakage, and the way to tell is to vary the perturbation and see whether the response follows.**

3.8.7 The norm this implies, and our own miss

The defect is invisible at $T \leq 257$. Our own census in §3.6 ran at T = 48. **Had we not questioned the test length, we would have published this model as clean.**

We state the resulting rule plainly, including as a correction to our own procedure:

> **The audit sequence length must exceed the model's chunk, window, and kernel parameters.** Otherwise the corresponding code paths degenerate and an entire class of defects is unobservable.

This joins the positive-control gate of §3.7.2 as the second methodological requirement we adopted only after our own instrument failed us.

4 Limitations

We list these in descending order of how much they should change a reader's confidence.

1. **Two released models, but one code lineage — not a base rate — and validated asymmetrically.** We found genuine causal violations in `Zyphra/Zamba2-1.2B` and `nvidia/Nemotron-H-8B-Base-8K` (§3.8) and diagnosed both to the same three lines. The evidence is now nearly symmetric. Zamba2 carries the original full battery (six alternative explanations ruled out, four seeds, a random-weight `from_config` structural test, an on-checkpoint patch to exact zero); Nemotron-H has since received **four-seed reproducibility, a live positive control on the loaded checkpoint, and its own on-checkpoint patch driving the leak to exact zero** (§3.8.5). The only pieces not re-run for Nemotron-H are the random-weight `from_config` build and the six-way alternative-explanation table; its structural character we take from the byte-identical code together with the on-checkpoint patch-to-zero result. Because the two inter-chunk blocks are byte-identical, the finding is best read as **one defect instantiated twice through shared code**, not two independent data points. It establishes that the property is violated in practice, in released models at 1B–8B scale, in the reference PyTorch path (§3.8.2); it is **not** enough to support any statement about how common such defects are across the ecosystem, and we make none. Treat it as an existence proof with a traced lineage, not a rate.
2. **We do not improve on existing practice at detection.** A logits-only perturbation check (B1) and a resampling variant (B4) detect every injected fault we detect. The contribution is layer localization. Readers who need only a yes/no answer already have a method, and it is one line shorter than ours.
3. **A gradient-based check matches our localization exactly.** B5 scores 96/96 on both detection and localization. Our case against it rests on cost (a backward pass and retained activations), applicability (differentiability of every op on the path), and the interpretive caveat that a zero gradient does not entail absence of dependence — not on accuracy.
4. **Our own public releases are only partially audited** (§3.7.4). The primary release is now audited under both controls — 49 layers clean, 16/16 positive-control localization. The instruction-tuned variant has a negative control only, and two further internal checkpoints remain unaudited. The earlier blanket load failure was a loading-path conflict, since diagnosed and resolved.
5. **The census is small and not a random sample.** Roughly twenty checkpoints in total, selected for architectural diversity rather than sampled, ranging from 129M to 9B parameters. Scale confers no protection in either direction: the largest clean model is Bamba-9B, while an 8B model (Nemotron-H) is one of the two leakers — so the earlier "all defects under 3B" caveat is removed, but not by showing large models are safe. Twenty non-randomly-chosen models still cannot support proportions or base rates, and we report none.
6. **Part of the ecosystem is unauditable on CPU** (§3.7.3). Four checkpoints require GPU-only compiled kernels; one is gated; one is not loadable through the standard interface. Coverage of custom-kernel hybrids is therefore systematically thinner than coverage of reference implementations — and custom-kernel implementations are plausibly where bugs are more likely, so this gap works against us.
7. **The positive-control gate does not cover every §3.8 clean verdict.** The gate (§3.2) is firmly satisfied for the from-config reference construction (clean *and* a live, propagating-fault-localizing positive control, §3.8.4), which is what carries the "conformant code is clean" claim. But applying it to each *released* conformant checkpoint requires injecting a fault at a layer's output, and several architectures' layer return signatures do not accept our injection (the shift has no effect, so the positive control cannot fire) — Falcon-H1 is one such case, where the model is verified input-responsive yet un-injectable in our harness. Those clean verdicts in Table 7b are therefore **provisional**: consistent with the prediction, but not gated. A per-architecture injection adapter would close this, and we did not build one.
8. **The exact-localization floor is model-dependent and one model is unexplained.** `LFM2-1.2B` localized all 24 $\varepsilon = 1.0$ injections exactly, yet localized none of the swept

points at $\varepsilon \leq$ 1e-1, consistently naming a layer three positions downstream. We report the anomaly without a mechanism.

9. **Limited configuration coverage — and one dimension of it proved decisive.** The injected-fault suite and Table 4 census ran at a single length (T = 48), one seed, one batch size, evaluation mode only. §3.8 shows this was not a harmless simplification: varying only the sequence length surfaced a defect that was otherwise invisible. We have since audited at chunk-exceeding lengths, but the remaining dimensions are still uncovered — faults that appear only for particular batch shapes, only under training-mode stochasticity, or only in distributed or sharded execution paths are outside what we tested. Given that one unexamined dimension already hid a real defect, we regard the others as open risk rather than as unlikely.
10. **Static, structural faults only.** We assume the fault is present in the graph for every input. Data-dependent leakage triggered by rare inputs would require a different search strategy.
11. **B6 was measured on one model.** The clean-model false positive (1.850e-04) is a single observation and should not be generalized to the practice as a whole without replication.
12. **The claim of no prior work is a search result, not a proof.** Our gap statement (§2) rests on keyword-level cross-referenced search in English.

5 Conclusion

Prefix invariance is the one property every autoregressive model must satisfy and essentially no release reports. The check costs two forward passes, needs no training, no gradients, no labels, and no accelerator, and it names the layer at fault rather than only raising a flag.

The check that the field actually performs — reading the attention mask — is unsound in principle, because causality is a property of the whole map and the mask governs one operator. Our measurements make the gap concrete: static mask inspection flagged **0 of 192** injected faults, and for two of the eight audited architectures it has no object to inspect at all. As stacks continue to mix masked and unmasked mixers, that coverage gap widens.

The check found real defects, and found them systematically. A static census of the chunked-scan implementations shipped in `transformers` predicted, before running anything, that two of six — `Zamba2` and `Nemotron-H` — would violate causality, because their inter-chunk recurrence reduces over the wrong axis. A dynamic audit confirmed both: in each, perturbing the final token contaminates earlier positions beginning exactly at the model's chunk boundary (256 and 128); the leak survives randomly initialized weights, and vanishes under a two-line correction. The four conformant implementations, and a from-config reference control, stay bit-exact clean. The two failing implementations carry a byte-identical error, so this is **one defect traced across two released models** — a lineage, not a rate — and we generalize no further than that. We reported the Zamba2 instance upstream and disclose the Nemotron-H instance in this paper. What is settled is that the property is violated in shipped models from major providers, and that a reader can predict where from the source alone.

We are equally careful about what we did not show. We do not detect better than a logits-only check that practitioners already use informally; we localize, and that is the delta. A gradient-based check localizes as well as we do, more expensively. Two further checkpoints produced small nonzero deltas that we deliberately declined to report as findings, because a perturbation sweep showed the response was flat in the perturbation magnitude.

Twice, our own instrument was the thing that had to be corrected. Three checkpoints returned Δ = 0 at every layer while their positive controls returned 0/24 — a false clean verdict averted only by requiring the instrument to demonstrate liveness on the artifact under test. And the defect above is invisible at the sequence length our own census used; had we not questioned that length, we would have published the model as clean.

Both cases generalize past this paper:

> **Any audit that concludes from an absence of signal owes its reader evidence that the instrument was live** — a positive control on the same loaded artifact. **Any audit of a hybrid stack owes its reader a test length exceeding the model's**

own chunk, window, and kernel parameters — otherwise the code paths that carry those defects are never exercised.

We propose one norm. **Release a causal-correctness certificate with every architecture release**: the per-layer maximum prefix deltas, the threshold, the precision, **the sequence length relative to the model's internal size parameters**, and the positive-control result on that same checkpoint. It belongs beside the parameter count — a number reported not because it is interesting when correct, but because its absence is the problem.

References

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Technical Report*, 2018.

[2] Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *NeurIPS*, 2014.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[7] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. In *ACL*, 2020.

[10] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. In *arXiv preprint arXiv:1904.10509*, 2019.

[11] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, et al. Big bird: Transformers for longer sequences. *NeurIPS*, 2020.

[12] Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *ICLR*, 2022.

[13] Bo Peng, Eric Alcaide, Quentin Anthony, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[14] Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 26670–26698. PMLR, 2023.

[15] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. *International Conference on Learning Representations (ICLR)*, 2022.

[16] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2024.

[17] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

[18] Soham De et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint*, 2024.

[19] Aleksandar Botev, Soham De, Samuel L. Smith, et al. Recurrentgemma: Moving past transformers for efficient open language models. *arXiv preprint arXiv:2404.07839*, 2024.

[20] Michael Poli et al. Hyena hierarchy: Towards larger convolutional language models. *ICML*, 2023.

[21] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 2022.

[22] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

[23] Xiaoyuan Xie, J. W. K. Ho, Chris Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, 84(4):544–558, 2011.

[24] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *SOSP*, 2017.

[25] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *ICSE*, 2018.

[26] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[28] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep neural network ensembles for time series classification. *arXiv preprint arXiv:1903.06602*, 2019.

[29] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *RecSys*, 2019.

[30] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems*, 39(2), 2021.

[31] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Re. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 2022.

Table 12: Executable injected-fault specifications grouped into four classes of temporal leakage. Each transformation is applied to a selected layer output and serves as a positive-control fault throughout the evaluation.

Class	Pattern	Transformation
C1	`shift`	`concat(o[:, 1:], o[:, -1:])` — output pulled one step earlier
C1	`peek1`	`o + eps * concat(o[:, 1:], o[:, -1:])`
C2	`pool_leak`	`o + eps * mean(o, dim=time, keepdim)`
C2	`global_max`	`o + eps * max(o, dim=time, keepdim)`
C2	`future_mean_k`	`o + eps * mean of the next k = 4 positions`
C3	`norm_leak`	`(1-eps)*o + eps*(o-mu)/sigma`, with μ, σ over `(time, hidden)` jointly
C4	`rev_scan`	`o + eps * flip(cumsum(flip(o,time),time),time)/T`
C4	`last_broadcast`	`o + eps * o[:, -1:, :]`

Appendix A — Reproduction

A.1 Why we print the method instead of shipping a package

The method is five lines of arithmetic on top of standard forward hooks. A competent reader can reimplement §2.2 in under an hour against any model whose layers are enumerable, and reimplementation is a better reproduction than running our binary, because it is independent.

We therefore release **measurements rather than a package**: complete audit logs, exact checkpoint identifiers, per-layer delta arrays for every clean scan and every injected trial, the injected-fault specifications, and the environment description below. Everything needed to reproduce or contest any number in this paper is in that set. No license is granted here to any implementation, and none of the released material is a software distribution.

A.2 Environment

CPU for the census and injection suite (Intel Xeon Gold 6526Y, 64 threads, 251 GB RAM); a single GPU for the $\geq$1B-parameter audits (§3.7.4, §3.8), for memory capacity only — the test needs no accelerator. float32 unless stated; float64 for §3.5. PyTorch 2.10.0, standard model-hub loading with remote code enabled where the architecture requires it, evaluation mode, caching disabled. Fused sequence-model kernels (`causal_conv1d`, `mamba_ssm`) absent, so state-space models ran reference implementations — this is the environment in which the §3.7.1 failure occurred and should be assumed to matter.

A.3 Injected-fault specifications

Each pattern wraps one layer's forward and transforms its output tensor $\mathbf{o}$ of shape (B, T, d). All are exactly reproducible from these definitions.

Injection depths: layer 1, $\lfloor L/2 \rfloor$, and $L - 2$. Injection strength $\varepsilon = 1.0$ except in the sweep of Section 3.5.

A.4 Baseline specifications

- **B1 logits-only:** the same two inputs; compare final logits over positions $[0, T-1)$; flag if max difference $> \tau$. No layer information available by construction.
- **B2 mask inspection:** enumerate modules; flag any exposing a causal/`is_causal` attribute set false. Static; zero forward passes.
- **B3 shuffled-suffix perplexity:** shuffle input positions $\geq$ T/2; compare cross-entropy on targets ¡ T/2; threshold 1e-4.
- **B4 future-token resampling:** resample the final token with 3 seeds; flag if any produces a logit change over the prefix; 4 forward passes.
- **B5 gradient:** backpropagate the summed prefix logits and inspect gradients at future positions of each layer's output; the deepest layer with nonzero future-position gradient, plus one, is the reported layer. 1 forward + 1 backward, activation graph retained.

Table 13: Primary experimental scripts used throughout the evaluation. The table summarizes the main audit harnesses, baseline implementations, and sensitivity-analysis programs used to generate the reported results.

File	Role
`leak_cpu_suite2.py`	Census harness: detector + injections + B1–B3; per-model JSON flush; records failures with traceback
`baseline_e2.py`	Extended comparison: detector + B1–B6, bit-level equality check, ε sweep
`fp64_sens.py`	Precision/threshold sweep (§3.5)

Table 14: Result files associated with the public-checkpoint census (§3.6) and the attempted-but-not-audited cases summarized in Tables 4 and 5. The table maps released JSON artifacts to the models, audits, and failure analyses discussed in the paper.

File	Contents
`results/smoke_smol.json`	SmolLM2-135M census record
`results/tier1.json`, `results/tier1__*.json`	Qwen3-0.6B, mamba-130m-hf
`results/b3__LiquidAI_LFM2-1.2B.json`	LFM2-1.2B, including the §3.5 floor anomaly (`sensitivity.sweep`)
`results/b5__RWKV_rwkv-6-world-1b6.json`	RWKV-6 1.6B
`results/b5__google_gemma-3-1b-it.json`	Gemma-3-1B-it
`results/b5__google_recurrentgemma-2b.json`	RecurrentGemma-2B
`results/b1__tiiuae_Falcon-H1-0.5B-Base.json`, `b1__...-1.5B-Instruct.json`, `b3__...-7B-Base.json`	§3.7.1 — clean $\Delta = 0$ with positive control 0/24
`results/b1__nvidia_Hymba-1.5B-Base.json`, `b2__microsoft_Phi-4-mini-flash-reasoning.json`, `b2__state-spaces_mamba2-2.7b.json`, `b2__RWKV_rwkv-6-world-1b6.json`, `b3__togethercomputer_StripedHyena-Nous-7B.json`, `b3__Zyphra_Zamba2-7B.json`, `b3__google_recurrentgemma-2b.json`, `b5__Zyphra_Zamba2-7B.json`	Load failures (Table 5). `b2__RWKV...` and `b3__google_recurrentgemma...` are earlier failed attempts later resolved in `b5`.
`results/aether_self.json`, `results/aether_self__*.json` ($\times 4$)	§3.7.4 — our own releases, four load failures

- **B6 cache consistency:** compare incremental cached decoding against a single full-sequence forward; threshold 1e-4; T+1 forward passes.

A.5 Known defect in a superseded artifact

An early version of the sensitivity harness cast hidden states to float32 before computing deltas, truncating float64 measurements to float32 resolution (visible as deltas quantized to powers of two). It was corrected to compute in float64, and §3.5 uses only post-correction data. The superseded file is retained and marked in Appendix B so that the correction is auditable. float32 results were unaffected (the cast was lossless there) and were re-verified.

Appendix B — Raw data index

All paths are on the compute host under `/data/ginipick/aether_final/N-paper-exp/cpu/`. Every table and number in §3 is traceable to a file below. Retained per-record fields include the full per-layer delta array for every clean scan and every injected trial, wall time, and peak RSS.

Table 13 summarizes the primary scripts used to generate the reported results, including the census harness, baseline implementations, and precision-sensitivity analyses.

Table 14 lists the released result files associated with the public-checkpoint census, including successful audits, load-failure records, and internal validation runs.

Table 15: Result files associated with the baseline comparison (§3.3, Table 1) and bit-level verification experiments (§3.4, Table 2). The table maps released JSON artifacts to the corresponding evaluations and checkpoints.

File	Contents
`results/e2_smol_fp32.json`	SmolLM2-135M: all six baselines including B6; `clean_bitwise`; clean-model B6 false positive `1.850128173828125e-04`
`results/e2_multi_nob6.json`	Qwen3-0.6B, pythia-1.4b, mamba-130m-hf: B1–B5; `clean_bitwise` for each

Table 16: Result files associated with the precision and threshold sensitivity analysis (§3.5, Table 3). The table lists the released JSON artifacts used to characterize exact-localization floors under different numerical precisions and threshold settings.

File	Contents
`results/sens_smol_fp32.json`	float32, $\tau = 10^{-6}$, 45 points (3 depths $\times$ 15 ε)
`results/sens_smol_fp64_v2.json`	float64, $\tau = 10^{-6}$, 45 points
`results/sens_smol_fp32_tau1e-14.json`	float32, $\tau = 10^{-14}$
`results/sens_smol_fp64_tau1e-14.json`	float64, $\tau = 10^{-14}$
`results/sens_smol_fp64.json`	**SUPERSEDED — do not cite.** Contains the Appendix A.5 float32-cast defect

Table 15 summarizes the released JSON artifacts underlying the baseline-comparison and bit-level-verification experiments.

Table 16 summarizes the released artifacts underlying the precision and threshold sensitivity study reported in Table 3.