Neural Distinguishers Expire on Carry Composition

Peter Hollows

June 22, 2026

Abstract

A hand-built carry-aware score measures a one-round advantage cliff on reduced SHA-256 [2]: it predicts a downstream output byte one adder layer away and loses all reach one round deeper. The natural objection is that a learned model might find local structure a human feature misses, as Gohr's neural distinguishers did for round-reduced Speck [3]. We translate that methodology to the mining read point. A residual network receives, as features, everything computed through an interior round (the state words, their carry-free derivations, the round's modular sums, and the schedule words, in bit, value, and Fourier encodings) and is free to learn any function of them. The network exceeds the hand-built score one adder layer downstream (0.998 versus 0.88 retained advantage) and then collapses to the noise floor one round deeper, across independent stems, a shifted read point, and a 90× capacity, 64× data, and 8× training-time scaling. On pure k-operand modular sums, with no SHA structure present, the same network learns the top byte for $k \leq 3$ and fails for $k \geq 4$. We show the wall is not an artifact of finite feature precision (a seed-paired `float32`-versus-`float64` comparison is null in both arms) and that it is the learner's reach rather than an intrinsic boundary: the carry chain's spectral gap is $\frac{1}{2}$ for every k [2], so nothing in the carry's mixing singles out $k = 4$. We also identify a distinct second failure mode, feature isolation, and show it does not affect the main result.

1 Introduction

The first paper in this series [1] reduces the question of whether mining beats brute force to a single residual: a global algebraic shortcut that resolves the output without resolving the carries on the path. The second [2] measures how far the natural local methods reach toward such a shortcut, using a hand-built carry-aware score, and finds a one-round cliff: strong selection of a downstream output byte one adder layer away, below detection one round deeper, against a 386-layer separation.

A single hand-built feature invites one objection above all. A learned model, searching the local feature space with gradient descent, might find structure the human score discards. This is precisely what Gohr did to round-reduced Speck [3], an ARX cipher with the same modular-addition nonlinearity: a trained distinguisher beat the state-of-the-art hand-built differential distinguisher. We run that experiment at the mining read point and report four findings. The learned attacker beats the hand score where signal exists and dies at the same single-round boundary (Sections 2–4). On pure carry composition, free of any SHA structure, it expires between two and three chained additions (Section 5). The wall survives full double-precision rebuilding (Section 6) and is the learner's reach rather than an intrinsic property of the arithmetic (Section 7). A separate failure mode, the inability to pull relevant bits out of distractors, is real but does not drive the headline (Section 8).

2 Setup

We use the selection-advantage metric of the anchor sweep [2]: take the best fraction 1/256 of candidates by predicted target, and report $1-\overline{y}_{\text{selected}}/\overline{y}_{\text{all}}$, where y is the top byte of the downstream state word. Advantage 0 is chance; advantage near 1 is near-perfect selection. The metric is identical to the hand-built measurement, so the two attackers are directly comparable at each prediction depth j (rounds past the read point).

The network is a GELU residual multilayer perceptron of about 7.4 million parameters with a 256-way top-byte head, trained with cross-entropy on 2^{20} candidates and tested on 2^{19} at each cell, over 50 epochs. At an interior read round it receives, as engineered features, everything computed through that round: the state words, their carry-free derivations, the round's modular sums T_1 and T_2, and the schedule words, in bit, value, and byte-harmonic Fourier encodings. The protocol was pre-registered, with amendments recorded as the positive control was brought to pass.

That positive control is itself informative. The network learns the top byte of one two-operand modular addition, topbyte$(T_1 + T_2)$ from Fourier-featured operands, to advantage 0.998. It failed, at every capacity, feature, and optimizer setting tried, to compose T_1 and T_2 from their seven constituent words, a depth of about five adder layers, until those sums were supplied directly. The learned attacker's demonstrated arithmetic reach is therefore between one and about five adder layers, which bounds how the later nulls should be read.

3 The one-round cliff

Read at an interior round, with the modular sums supplied so the network starts one adder layer from the target, the learned distinguisher reaches advantage 0.9978 (range 0.9972 to 0.9990 over three stems), against the hand-built score's 0.8841. The network learns the carry-in corrections that the truncated $(T_1 \gg 24) + (T_2 \gg 24)$ score discards, so it is a strictly stronger member of the same local attack class.

One round deeper, the advantage is gone. Every $j \geq 1$ cell sits at the noise floor. The unbiased final-epoch statistic, pooled over the twelve $j \geq 1$ cells, is -0.0040 with 95% confidence interval $[-0.0079, -0.0001]$; at $j = 1$ alone it is -0.0060, interval $[-0.0126, +0.0007]$. A configuration-matched shuffled-label control reads 0.0212, inside the $j \geq 1$ band, which pins those readings as the pipeline's maximum-over-epochs noise floor rather than residual signal. The learned attacker, stronger than the hand score where signal exists, retains no advantage one round past the read point. The pattern holds at a shifted read point and across independent stems.

4 Scaling does not move the wall

A null at $j \geq 1$ could in principle be a capacity or training limit rather than a reach limit. We scaled each axis independently at $j = 1$, pairing every cell against a configuration-matched shuffle-label run, with the verdict being the gap between the real and shuffle maxima (since maximum-over-epochs advantage inflates with capacity and epoch count, a fixed gate would manufacture signal as scale grows). Capacity from 0.4 to 35.7 million parameters (90×), data from 2^{16} to 2^{22} samples (64×), and training time from 50 to 400 epochs (8×, a grokking probe [4]) all return null, with no dose-response on any axis and no late generalization transition. The $j = 1$ null is a reach limit: a 90× capacity range, a 64× data range, and an 8× time range do not move the learned attacker past one round. (An earlier 500× capacity endpoint was discarded as invalid, since its test predictions collapsed to a constant from epoch 35 on; the defensible range is 90×.)

5 Where learning stops: the k-operand ladder

The $j \geq 1$ nulls confound two readings: that SHA's carry structure destroys the signal, and that this optimizer cannot compose modular arithmetic in this format. We separate them on data with no SHA structure at all. The task is the top byte of a sum of k uniform 32-bit operands, with the same pipeline and features. The network solves $k = 2$ (0.9996) and $k = 3$ (0.9983), then fails at $k = 4$ and through $k = 7$ (maxima 0.025 to 0.041, final-epoch advantages near zero). The learned attacker expires on carry composition itself, between two and three chained additions, on synthetic data.

operands k	2	3	4	5–7
max advantage	0.9996	0.9983	0.041	dead (≤ 0.04)

A finer probe sharpens the death point. Holding $k = 4$ and sweeping the operand width, the task is learnable at widths 8 through 28 bits and dead at width 32, with training bimodal below full width: a seed either solves the task almost perfectly or sits at the floor, with little in between, and at width 32 no seed of many has left the floor. The $k = 4$ death is therefore a joint condition, $k \geq 4$ and full width, and it has the character of an optimization cliff whose success probability falls to zero rather than a smooth capacity limit.

operand width (bits)	8	12	16	20	24	28	32
seeds reaching $\sim$1	1/2	2/2	1/2	1/2	3/3	1/2	**0/8**

6 The wall survives double precision

The bimodal width result admits a precision reading: the value and Fourier features resolve an operand to about 2^{-24} in single precision, so an analog shortcut (estimate the real-valued sum, read its top byte) would lose the carry-relevant low bits exactly when the operand width exceeds the mantissa, near width 32. We tested this directly. For each seed we built the $k = 4$, width-32 task twice, with identical integer operands and identical parameter initialization, differing only in numeric precision: full double precision, features and network arithmetic, against the single-precision baseline. (Full double precision is required, since any cast back to single precision after the features reintroduces the same 2^{-24} wall.)

The two arms are null together. Across six seed-paired runs, both precisions leave the advantage at the floor, 0/6 seeds learning in either, with the double-precision arm tracking the single-precision arm to within about 0.01 at every seed and a maximum advantage of 0.086. Restoring full operand resolution does not let gradient descent find the carry circuit. On this evidence the wall is a trainability barrier, not an artifact of finite feature precision.

7 The learner's reach

The $k = 4$ threshold is a property of the learner, not of the carry chain. The carry chain for k operands has contraction rate $\frac{1}{2}$ for every k [2]: its eigenvalues are $\{2^{-j}\}$, and larger k adds only faster-decaying modes, so nothing in the carry's own mixing distinguishes $k = 4$ from any other operand count. The intrinsic correlation decay is set by the base, not by k. What the ladder isolates is the reach of gradient descent over fixed features at composing carries, a statement about this attack class.

This is the same conclusion the precision result reaches by a second route. Together they say the $k = 4$ cliff is where the gradient-descent attacker's last analog shortcut runs out, not an information

boundary in the arithmetic. The carry, the one part of modular addition with no low-degree or analog surrogate, is what the learner cannot synthesize; this is consistent with Gohr's distinguishers, which exploited statistical bias already present in the data they were given rather than synthesizing cipher arithmetic.

8 A second failure mode: feature isolation

Probing the ladder surfaced a separate way the learner fails, which we record so it is not confused with carry depth. Holding $k = 4$ and full width and varying which byte of the sum is predicted, every read-out is dead, including the low byte. The low byte of the sum is, mathematically, an 8-bit modular addition of the operands' low bytes, the same easy task a width-8 cell learns, so its death is not about carry depth. Varying the feature set isolates the cause: with all features, or with all 32 bits and no analog features, the low byte stays dead; with only the eight relevant low bits, a seed revives to advantage 0.45. The killer is the presence of 24 distractor bits, not the analog features. The learner cannot extract the relevant sub-circuit when its inputs are buried among irrelevant ones.

This is a weakness of a multilayer perceptron over a fixed, wide feature set, in principle addressable by attention or feature selection, so it is not an intrinsic-hardness claim. It does not touch the main result: the k-operand ladder and the width ladder use top-byte targets, and the top byte of a k-operand sum depends on every operand bit through the carry chain, so no input bit is a distractor there. A Gohr-faithful convolutional architecture over bit positions was also run and reached only 0.03 to 0.06 on a task the perceptron passes above 0.99; its inductive bias is matched to the XOR-differential locality Gohr's distinguishers exploit, which carry chains do not have.

9 Scope, reproduction, and conclusion

This tests a one-shot learned predictor, the direct analogue of Gohr's distinguisher. It does not rule out machine learning used as a search heuristic over algebraic representations; Gohr's strongest attacks combined the network with classical key-ranking and multi-round search, which is the same residual the first paper [1] scopes out and does not claim to close. The experiment uses the simplified single-compression read point of the anchor sweep [2], faithful to the round-function mechanism but not the full mining map, and "null" means below a pre-registered gate at the tested power, consistent with zero rather than a proof of zero.

The experiment is reproducible from a pre-registered protocol with documented amendments, training code, and raw per-cell results; unlike the two static results of [2] it requires a GPU and the JAX/optax stack. The vendored SHA-256 is verified bit-exact against `hashlib`.

A learned distinguisher of the kind that broke round-reduced Speck beats the hand-built carry-aware score one adder layer downstream and retains no advantage one round deeper. It expires on carry composition between two and three chained additions, a wall that is independent of feature precision and is the reach of the learner rather than a boundary in the arithmetic. The one-round cliff of [2] is not an artifact of a weak hand-built score, and the 386-layer separation stands against the strongest local attack we could train.

References

[1] P. Hollows. *The Amortization Envelope: Where a SHA-256d Mining Advantage Can and Cannot Live.* Carry-Adder Wall series I, 2026.

[2] P. Hollows. *Carry Depth: A Circuit-Independent Coordinate of ARX, and the One-Round Advantage Cliff.* Carry-Adder Wall series II, 2026.

[3] A. Gohr. *Improving Attacks on Round-Reduced Speck32/64 Using Deep Learning.* CRYPTO 2019.

[4] A. Power, Y. Burda, H. Edwards, I. Babuschkin, V. Misra. *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets.* 2022, `arXiv:2201.02177`.