Breaking the Beyond-Birthday-Bound Security of ♯Pencil

Léonard Assouline* and Cécile Delerablée†

Leanear

August 11, 2026

Abstract

♯Pencil is a domain-extended pseudorandom function by Bhaumik *et al.*, accepted at CRYPTO 2026, claiming that it achieves close to n-bit security beyond the birthday bound. It is used as the key-derivation layer of the ♯Pencil-CAU authenticated-encryption mode. We show that ♯Pencil has a birthday-bound collision attack: its front end Sharp compresses the second half N_2 of the input through the $(n-8)$-bit value

$$J(N_2) = \mathrm{msb}_{n-8}\big(\mathsf{E}_{K_1}(N_2 \,\|\, \texttt{0x00})\big),$$

after which the entire computation is a deterministic function of (N_1, J). Thus, for any fixed N_1, distinct values N_2, N_2' satisfying $J(N_2) = J(N_2')$ produce identical ♯Pencil outputs. Such collisions occur with probability $1 - e^{-1}$ after $2^{(n-7)/2}$ queries, $2^{60.5}$ when $n = 128$. This yields a PRF distinguisher with constant advantage, contradicting the security bound of Theorem 4. Since $2^{60.5}$ falls below the birthday bound $2^{n/2}$ that the construction was designed to pass, the beyond-birthday-bound property does not hold. Given the derived key K_1, an explicit collision can be constructed in approximately 10^3 inverse-cipher evaluations in expectation. The same collision breaks ♯Pencil-CAU, as two nonce-respecting queries can reuse the key and the nonce of the inner GCM instance, yielding the difference of the two plaintexts. The first version of the paper does not have the defect: it left J untruncated, which makes the map injective and invalidates the collisions above. Reverting to it is the remedy we suggest.

1 Introduction

In the Galois/Counter Mode (GCM) of operation for AES, when the nonce is generated at random, the 96-bit nonce allows only 2^{32} invocations per key before the risk of collision becomes too high. Moreover, a single reused nonce breaks a key's security. To address this, NIST published a pre-draft call for comments [Nat25], which suggested among other things a key-derivation function wrapping GCM and capable of absorbing a long nonce and handing the mode a fresh (key, nonce) pair.

This last call was then obsoleted on June 1st 2026 [Nat26], however ♯Pencil, by Bhaumik, Degabriele and Dhar [BDD25] was proposed in response to this request. ♯Pencil is a domain-extended PRF taking a $(2n-16)$-bit input to a $5n$-bit output. It is constructed as the composition of a three-call preprocessing layer Sharp with an augmented XORP layer Pencil. The composition ♯Pencil-CAU admits an AEAD security proof, and security is claimed not to degrade to that of plain GCM when nonces are chosen adversarially. Theorem 4 of [BDD25] claims $\mathsf{Adv}^{\mathrm{PRF}}_{\sharp\mathsf{Pencil}}(\mathcal{A}) \le q/2^{n-15} + nq/2^{n-4} + 3p/2^{2n} + 6/2^n$, i.e. roughly $q/2^{113}$ at $n = 128$: security close to the full n bits, and well beyond the birthday bound.

*la@leanear.io

†cd@leanear.io

We show that ♯Pencil does not meet this bound, and perform a birthday attack where collisions appear after $\Theta(2^{(n-7)/2})$ queries, about $2^{60.5}$ at $n = 128$, yielding a PRF distinguisher of constant advantage. The construction exists in order to pass the birthday bound $2^{n/2}$, and $2^{60.5}$ falls below it, so the beyond-birthday-bound claim fails, and with it the nonce-respecting AEAD column of [BDD25, Fig. 2]. We locate the error in the security proof: the claim, made in the proof of Lemma 7 and reused in those of Lemmas 8 and 9, that the map $\mathsf{N} \mapsto \mathsf{W}$ is injective. The remedy is a revert: the first version of [BDD25], [BD25], separated the front end by key rather than by input bytes, thus it did not truncate J, which is what made $\mathsf{N} \mapsto \mathsf{W}$ injective.

2 The Sharp function

Let E be the block cipher (n-bit block, $2n$-bit key, instantiated with AES-256, so $n = 128$), and K_1 a subkey derived from the master key K. On input $N = (N_1, N_2)$ with $|N_1| = |N_2| = n-8$, Sharp computes [BDD25, Fig. 5]

$$J \;=\; \mathrm{msb}_{n-8}\big(\mathsf{E}_{K_1}(N_2 \,\|\, \texttt{0x00})\big), \qquad W_1 = N_1 \oplus J, \qquad W_2 = N_1 \oplus 2{\cdot}J,$$

followed by $R_1 = \mathrm{msb}_{n-8}(\mathsf{E}_{K_1}(W_1 \| \langle 1 \rangle_8))$ and $R_2 = \mathrm{msb}_{n-8}(\mathsf{E}_{K_1}(W_2 \| \langle 2 \rangle_8))$, ♯Pencil then outputs $\mathsf{Pencil}_{K_2}(W, R)$, a deterministic function of (W_1, W_2, R_1, R_2).

The map $N_2 \mapsto J$ is a permutation followed by an 8-bit truncation, it is thus not injective with overwhelming probability over the choice of K_1. The reason why the permutation property does not protect it is the subject of the remark in Section 3. In Section 4, we show how this contradicts the proof of Lemma 7 in Appendix C.4, Lemma 8 in Appendix C.5 and of Lemma 9 in Appendix C.6 of [BDD25], upon which the central theorem is built.

3 The collision

Theorem 1. *♯Pencil is not collision resistant: collisions are found below the birthday bound, so it is not a beyond-birthday-bound PRF and the bound of Theorem 4 of [BDD25] does not hold.*

The rest of the section proves Theorem 1.

Observation. *Every output of ♯Pencil is a function of (N_1, J) alone. Therefore, for any fixed N_1,*

$$J(N_2) = J(N_2') \implies \sharp\mathit{Pencil}_K(N_1, N_2) = \sharp\mathit{Pencil}_K(N_1, N_2').$$

Proof. W_1, W_2 are functions of (N_1, J) by definition, R_1, R_2 are functions of (W_1, W_2), and $\mathsf{Pencil}_{K_2}(W, R)$ is a function of (W, R). With N_1 fixed the entire chain is a function of J. □

Remark (Collisions in J are common). Although E_{K_1} is a permutation and no two inputs share a full n-bit image, truncating 8 bits cannot be done safely. Let us look at how many N_2 share a value of J. Each $(n-8)$-bit value of J labels a set of 2^8 full images, whose preimages are spread over all of $\{0,1\}^n$, and only those ending in 0x00 are reachable as $N_2 \| \texttt{0x00}$. The number of N_2 mapping to a given J is therefore Binomial$(2^8, 2^{-8})$, which is Poisson(1) up to $O(2^{-8})$. So $1 - 2e^{-1} \approx 26.4\%$ of the values of J have two preimages or more, a uniformly chosen N_2 has a collision in J with another N_2 with probability $1 - e^{-1} \approx 63.2\%$, and about 2^{n-9} colliding pairs exist.

Since J is $(n-8)$ bits, $J(N_2) = J(N_2')$ is a truncated-permutation birthday. For fixed N_1 and q distinct N_2's, let X count colliding pairs. We have:

$$\mathbb{E}[X] = \binom{q}{2} \cdot \frac{2^8 - 1}{2^n - 1} \;\approx\; \frac{q^2}{2^{n-7}} \;=:\; \lambda.$$

The probability of there being no collision is $\Pr[X = 0] = \frac{B(B-1)\ldots(B-q+1)m^q}{M(M-1)\ldots(M-q+1)}$ where $B = 2^{n-8}$, $m = 2^8$, and $M = 2^n = Bm$. Notice $\log \Pr[X = 0] = \log \prod_i \frac{M-im}{M-i} \approx -\sum_i \frac{i(m-1)}{M} \approx -\lambda$.

Thus, for $q = \Theta(2^{(n-7)/2})$, $\Pr[X = 0] = \exp(-(1 + o(1))\lambda)$, so with $q \approx 2^{(n-7)/2} = 2^{60.5}$ the collision probability approaches $1 - e^{-1}$, while Theorem 4 gives advantage below $2^{-52.4}$. With $q = 2^{62}$, $\lambda = 8$, the probability of a collision is almost certain ($1 - e^{-8} = 0.9996$). The side conditions of Theorem 4 hold here: $nq \le 2^{n-11}$ ($2^{67.5} \le 2^{117}$) and $25n^3 + 5n \le 2^{n/2}$ ($2^{25.6} \le 2^{64}$), and $p = 0$. Therefore Theorem 4 is false.

The distinguisher: $2^{60.5}$ queries, no key The birthday distinguisher is simple: given an instance of the PRF, fix N_1, then compute $q = 2^{60.5}$ evaluations of the PRF on distinct N_2, output 1 if any two outputs repeat. Against a uniform random function from $\{0,1\}^{2n-16}$ to $\{0,1\}^{5n}$ the same test returns 1 with probability $< 2^{-500}$ at $q = 2^{62}$, so the advantage of the distinguisher is almost exactly the collision probability computed above. As stated it stores q outputs. It can be made memoryless by iterating $N_2 \leftarrow \mathrm{msb}_{n-8}(Z_1)$ with cycle detection: a collision in that iteration is a collision in J about half the time, the rest being collisions in Z_1 itself, and comparing the two full outputs tells which.

An explicit collision: 10^3 operations, given K_1 This one is not an attack on the PRF game, since it uses a derived key; it is how the example below was produced. We can use the following procedure: sample a candidate value p for J and invert the 2^8 blocks $p\|c$. Then, the preimages of the form $(\cdot \,\|\, \texttt{0x00})$ are exactly the N_2 with $J(N_2) = p$.

```
find_collision(K1):
  repeat:
    p <- random (n-8)-bit string
    S <- empty list
    for c in 0..255:
      x <- E_K1^{-1}(p || c)
      if lsb_8(x) == 0x00:
        append msb_{n-8}(x) to S
    if |S| >= 2:
      return S[0], S[1]
```

Each iteration costs 2^8 inverse-cipher calls and succeeds with probability $\approx 1 - 2/e \approx 0.26$ (the count $|S|$ is approximately Poisson(1)), so the expected cost is about 10^3 inverse-cipher evaluations.

Explicit collision (AES-256).

```
K  = 000102030405060708090a0b0c0d0e0f          (the 32 bytes 0x00..0x1f)
     101112131415161718191a1b1c1d1e1f
K1  = E_K(<0>_128) || E_K(<1>_128)
    = f29000b62a499fd0a9f39a6add2e7780
      f05d76ae4ab99fe5a6f69b3148c2363d
K2  = E_K(<2>_128) || E_K(<3>_128)
    = 0ebcb5deb52c83bd08a8a935182c9199
      d24356532881602f809eb383c5ff5d56
N1  = 0^120

N2  = 5ec44e5641eaf6320ea2eb4273062a
N2' = 78d8cf9200b0cd79d0764956301bad

E_K1(N2 ||00) = d694690ca7f650150ced8b5e81892a0c
E_K1(N2'||00) = d694690ca7f650150ced8b5e81892a0d
```

```
both nonces then give the same intermediates:
J = W1 = d694690ca7f650150ced8b5e81892a
W2     = ad28d2194feca02a19db16bd031254
R1     = 8cc57bf1f87ee3257f500ffd6766db
R2     = 6a25b773a912e33cf598527e1487b8

#Pencil_K(N1||N2) = #Pencil_K(N1||N2') =
 9f9f8c0cecfecfd9387dad0cb61303a0f2fa723813e5f16759808b9c50e31165
 111017d8f6e4164152e2df59a56b15682289960589980ef184365934afe265ca
 218be8ffa3a203b3097389906ced7d4e
```

4 The error in the proof

The security proof of [BDD25] tracks, query by query, whether a new query brings fresh randomness into the analysis. A query is called *fresh* when at least one of its two values W_1, W_2 has not been seen before (i.e. it guarantees that R_1 or R_2 is a new E_{K_1} output), and *stale* when both have. The proofs of Appendices C.4 to C.6 are built on the assumption that stale queries never occur in two-query chains. The proofs of Lemmas 7 and 8 consider that a stale second query (where $\mathsf{W}^{(i_1)} = \mathsf{W}^{(i_2)}$) "is impossible because $\mathsf{N}^{(i_1)} \neq \mathsf{N}^{(i_2)}$, and the mapping $\mathsf{N}^{(i)} \mapsto \mathsf{W}^{(i)}$ is injective". As we have seen in Section 2, this is false.

The reasoning would be sound if the input to the PRF was (N_1, J). But the input is (N_1, N_2), with J obtained from N_2 by encrypting and then discarding eight bits, opening the door for a collision in J. When $N_1 = N_1'$, a collision in W_1 is exactly a collision in J, and it forces one in W_2 as well. The proof of Lemma 5 gives a probability of $(2^8 - 1)/(2^n - 1)$, or 2^{-120} at $n = 128$, whereas Appendix C.4 assigns probability zero to this event. Lemma 9 (Appendix C.6) is wrong as well: the bound of $2^8/(2^n - 1)$ assumes that one of W_1, W_2 still differs, but when both coincide so do R_1, R_2, all six X_j agree, and E_2 is certain. The rest of the attack follows from the birthday bound.

Notice that a collision puts the transcript among the bad ones of [BDD25, Sec. 3.2], whose probability Lemma 3 bounds by $2^{-52.4}$ at $q = 2^{60.5}$. We reach $1 - e^{-1}$, thus Lemma 3 is false by a factor of $2^{51.7}$.

4.1 A corrected bound

Repairing Lemma 3 costs one union-bound term. A collision on both coordinates of W requires $N_1 = N_1'$ together with $J = J'$, so

$$\Pr\left[\exists\, i \neq i' :\ \mathsf{W}^{(i)} = \mathsf{W}^{(i')}\right] \ \leq\ \binom{q}{2}\frac{2^8 - 1}{2^n - 1} \ \leq\ \frac{q^2}{2^{n-7}},$$

and Theorem 4 becomes

$$\mathsf{Adv}^{\mathrm{PRF}}_{\sharp\mathsf{Pencil}}(\mathcal{A}) \ \leq\ \frac{q^2}{2^{n-7}} + \frac{q}{2^{n-15}} + \frac{nq}{2^{n-4}} + \frac{3p}{2^{2n}} + \frac{6}{2^n},$$

where the distinguisher above shows the new term tight up to a factor of two. The security level is therefore $(n - 7)/2 \approx 60.5$ bits and not $n - 15 \approx 113$. Theorem 5 inherits the term through the game hop $G_2 \to G_3$, so $\sharp$Pencil-CAU inherits it.

Appendix F of [BDD25] rejects KC-XAES because its commitment tag collides at the birthday bound for an advantage of order $q^2/2^n$. Our attack gives an advantage of order $q^2/2^{n-7}$, the 8 bits of separation being what moves the exponent. The criterion used to disqualify KC-XAES therefore disqualifies $\sharp$Pencil. Note that Pencil is not the source of the break, the flaw comes from Sharp.

5 Conclusion

♯Pencil does not meet the bound of Theorem 4 of [BDD25]: collisions appear after $2^{60.5}$ queries, below the birthday bound $2^{n/2}$ that the construction exists to pass. We close with the consequence for ♯Pencil-CAU, and with the repair we suggest.

5.1 AEAD consequence

Two distinct nonces colliding in J as in Section 3 give identical ♯Pencil outputs, hence identical $(\widetilde{K}, \widetilde{N}, \mathsf{KC})$, so a fixed (AD, M) encrypts to the same ciphertext under both. This happens with probability $1 - e^{-1}$ for $q \approx 2^{(n-7)/2}$, against the probability of order 2^{-52} that Corollary 1 allows.

The consequence is stronger than a distinguisher. Encrypting M under N and M' under N' runs the inner CAU instance twice on the same $(\widetilde{K}, \widetilde{N})$, so the two keystreams coincide and $C \oplus C' = M \oplus M'$ on the common length. The two tags then yield a small set of candidate GHASH keys by the forbidden attack of [Jou06], so the proposed AEAD is broken for such nonces.

5.2 Repairing ♯Pencil

The proof needs $\mathsf{N} \mapsto \mathsf{W}$ to be injective, and a truncated encryption block does not give that. The first version of [BDD25], [BD25], does not have the defect: its front end, called Sharpen there, separates its three calls by key,

$$\mathsf{Sharpen}_{K_1,K_2,K_3}(N_1, N_2) \;=\; \big(\mathrm{msb}_{n-8}\left(\mathsf{E}_{K_2}(N_1 \oplus \mathsf{E}_{K_1}(N_2))\right),\; \mathrm{msb}_{n-8}\left(\mathsf{E}_{K_3}(N_1 \oplus 2{\cdot}\mathsf{E}_{K_1}(N_2))\right)\big),$$

[BD25, Sec. 4], so $J = \mathsf{E}_{K_1}(N_2)$ is left untruncated. Collapsing K_1, K_2, K_3 into the single key of [BDD25, Fig. 5] reserves a byte of each cipher input for domain separation. This is the truncation of Section 2. Our suggestion is to undo the collapse (though we would need to modify Sharpen so that it returns W as well to build the commitment).

Injectivity now comes back: if $W_1 = W_1'$ and $W_2 = W_2'$ then $J \oplus J' = N_1 \oplus N_1' = 2{\cdot}(J \oplus J')$, hence $3{\cdot}(J \oplus J') = 0$ with $2{\cdot}$ the product in $\mathrm{GF}(2^n)$ there; 3 is invertible, so $J = J'$ and $N_1 = N_1'$, and $N_2 = N_2'$ since E_{K_1} is a permutation of $\{0,1\}^n$.

References

[BD25] Ritam Bhaumik and Jean Paul Degabriele. Pencil: A domain-extended PRF with full n-bit security for strengthening GCM and more. Cryptology ePrint Archive, Paper 2025/383, version 20250228:113934, 2025.

[BDD25] Ritam Bhaumik, Jean Paul Degabriele, and Chandranan Dhar. ♯Pencil: A domain-extended committing BBB PRF for strengthening GCM. Cryptology ePrint Archive, Paper 2025/383, 2025.

[Jou06] Antoine Joux. Authentication failures in NIST version of GCM. 2006.

[Nat25] National Institute of Standards and Technology. Pre-draft call for comments: GCM and GMAC block cipher modes of operation. NIST Special Publication 800-38D Rev. 1 (Initial Preliminary Draft), January 2025. `https://csrc.nist.gov/pubs/sp/800/38/d/r1/iprd`.

[Nat26] National Institute of Standards and Technology. Second pre-draft call for comments: GCM and GMAC block cipher modes of operation. NIST Special Publication 800-38D Rev. 1 (2nd Preliminary Draft), June 2026. `https://csrc.nist.gov/pubs/sp/800/38/d/r1/2prd`.