THE FACTORIZATION WAS THE CIRCUIT: CHINESE REMAINDER THEOREM IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Emergent algorithms can arise in neural networks trained on simple tasks. Twenty small transformers (1-layer, no LayerNorm) trained to solve $a^b \bmod p$ over a range of primes p exhibit perfect generalization after grokking. We reverse engineer the emergent mechanism and find that the model first rewrites the answer as g^γ, and $\gamma = \log_g(a) \cdot b \bmod q$, where $q = p-1$ and g is a primitive root of p. It then splits that product using the *Chinese Remainder Theorem* (CRT): computing $\gamma \bmod m_i$ on independent streams, where m_i are the prime-power factors of q; then recombining the residues in the unembedding. Similar to prior work, we learn that the model uses a sparse set of "key" frequencies, which index the irreducible representations. Unlike prior work, this set is completely deterministic, as it is the union of disjoint bands of the Fourier spectrum at the multiples of q/m_i. We call these bands the CRT bands. Through causal tests, we show that each band controls its own corresponding stream independently. Furthermore, for prime-power streams $m_i = \ell^e$ (e.g., $\mathbb{Z}/16\mathbb{Z}$), the irreducible representations of these groups are filtered by order, dual to the quotient tower: $\mathbb{Z}/\ell^e\mathbb{Z} \twoheadrightarrow \cdots \twoheadrightarrow \mathbb{Z}/\ell\mathbb{Z}$. With the a priori knowledge of these irreducible representations and their frequencies, we are able to show the streams' independence causally, as well as isolate and observe the mechanism's circuitry formation during grokking.

1 INTRODUCTION

Grokking is the phenomenon in which a neural network suddenly generalizes after prolonged training; i.e., long after the network has achieved perfect prediction on the training set (Power et al., 2022). This phenomenon has become a central testbed for mechanistic interpretability: models trained on simple and specific algebraic tasks until they grok can be reverse-engineered. Nanda et al. (2023) laid the groundwork by reverse-engineering a transformer trained on modular addition. Their work has shown that the Discrete Fourier Transform (DFT) spectra of the model's components, such as the embedding and unembedding, are sparse; meaning that only a sparse set of "key" frequencies carries most of the spectral energy. In the language of Chughtai et al. (2023), the model implements the group's irreducible representations on the inputs. The multi-layer perceptron (MLP) implements modular addition on the inputs through a mechanism later dubbed the "Clock" mechanism by Zhong et al. (2023).

Subsequent work extended this study to other algebraic tasks in transformers. Furuta et al. (2024) reported that the DFT spectra of transformers trained on modular multiplication are dense; i.e., "multiplication requires cosine-biased components at all frequencies". Nguyen (2026) showed this density is an artifact of the analysis basis. Motivated by the group isomorphism between $(\mathbb{Z}/p\mathbb{Z})^*$ and $\mathbb{Z}/(p-1)\mathbb{Z}$, that work reindexed the token basis by the discrete logarithm. The result was sparse spectra, and the model indeed uses the Clock algorithm to perform addition over the reindexed inputs. Effectively, the model computes $a \cdot b$ through $\log(a) + \log(b)$.

This paper extends this line of work to modular exponentiation in transformers. We trained 20 1-layer transformers across prime moduli p and seeds until grokked. The generalization achieved by these models is consistent with the emergence of an algorithm during training, and we set out to reverse-engineer that algorithm. This paper shows that the mechanism can be interpreted as follows:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1. The models interpret the exponentiation $a^b \bmod p$ as $\alpha \cdot b \bmod q$ where $\alpha = \log_g(a)$, g is a primitive root of p, and $q = p - 1$.
2. The models decompose the product $\gamma = \alpha \cdot b \bmod q$ by the Chinese Remainder Theorem (CRT). Writing q as a product of pairwise coprime prime powers $q = \prod_{i=1}^{k} m_i$ with $m_i = \ell_i^{e_i}$, the theorem gives a ring isomorphism $\mathbb{Z}/q\mathbb{Z} \cong \prod_i \mathbb{Z}/m_i\mathbb{Z}$. The single product therefore splits into k per-component products $\gamma \bmod m_i$, each of which can be computed independently. We call these computation flows ***CRT streams***, or ***CRT components***; we use the two terms interchangeably.
3. Each stream occupies its own band of the Fourier spectrum, at the multiples of q/m_i. These bands are predetermined solely from the factorization of q.
4. The unembedding W_U recombines the components into the answer token g^γ by constructive interference across the bands, which is CRT reconstruction carried out in the read-out.

We then ask how this structure appears over training. Nanda et al. (2023) report that the grokking phenomenon consists of two phases: 1. The generalizing circuit forms before the test loss drops. 2. Weight decay removes the memorizing components. They show that the grokked model's DFT spectra of various components and circuits (e.g., W_U, W_E, W_{neur}) are sparse; and call the set of frequencies with outstanding energy "key frequencies". They show that these key frequencies index the irreducible representations of the group $\mathbb{Z}/p\mathbb{Z}$. The set of key frequencies is an indeterministic property of the trained model, as it differs from seed to seed even for the same model. In our task the corresponding set is not empirical. The CRT bands are fixed by the factorization of q; the frequencies in each band are at multiples of q/m_i, and are identical across seeds at a given p. Their union forms the set of key frequencies, and it is completely deterministic. With this knowledge, not only can we ablate the key frequencies of the model as a whole, but we can also surgically remove the frequencies of each CRT band individually to show stronger causal verification.

2 Related work

The Chinese Remainder Theorem in neural arithmetic. McCracken et al. (2025) propose that networks trained on modular addition implement an "approximate CRT". This means the network learns structures similar to the CRT decomposition to solve $a + b \bmod p$, and each neuron can be interpreted as a sinusoid on a coset or approximately so. For our work in particular, we show that the models implement the CRT decomposition *exactly*.

Modular exponentiation. Africa et al. (2025) study $a^b \bmod c$ with the modulus supplied as an input token, and reach 80.39% test accuracy aggregated over moduli. They report that several small moduli jump to near 100% together while the aggregate stays near 80%, but no case achieves perfect generalization. Although motivated by "modular exponentiation may present nontrivial number-theoretic patterns not present in simpler tasks", they have yet to find such patterns.

Nguyen (2026) introduces the discrete-logarithm reindexing this paper depends on, and reports the mechanism for modular *multiplication* as a Clock in the reindexed basis. It also reports a preliminary exponentiation result, a two-layer transformer at 100% on $a^b \bmod 41$, and states that the mechanism for exponentiation "remains under active investigation". This paper completes that investigation; this paper works on one-layer models instead.

Group-theoretic frameworks. Chughtai et al. (2023) showed that networks trained on group operations learn the group's irreducible representations, while Stander et al. (2024) found coset-based circuits instead on permutation groups. That disagreement is specific to the non-abelian setting, where cosets and irreducible representations compete as descriptions. Our exponent ring $\mathbb{Z}/q\mathbb{Z}$ is abelian, its irreducible representations are one-dimensional characters, and the factorization of its character group into prime-power characters is canonical.

Progress measures for grokking. Nanda et al. (2023) introduced restricted and excluded loss, both computed from a set of key frequencies, which is identified after the model has achieved grokking. Those frequencies carry no arithmetic label, so ablating one shows that a frequency matters but not which part of the answer it carries. Our bands are fixed by the factorization of q before training and each is labeled by the component it encodes. Therefore, we can remove a single band and state in advance which residue of the answer should fail.

3 THE CRT DECOMPOSITION

3.1 THE TWO MAPPINGS

The discrete logarithm $\log_g \colon ((\mathbb{Z}/p\mathbb{Z})^*, \times) \to (\mathbb{Z}/q\mathbb{Z}, +)$ is a group isomorphism, where p is a prime, $q = p - 1$, and g is any *primitive root* of p. This means:

$$g^q \equiv 1 \bmod p. \tag{1}$$

By writing $a = g^\alpha$, exponentiation turns into multiplication in the exponent: $a^b = g^{\alpha b}$. And since g has order q (Equation 1), the answer depends only on

$$\gamma \;=\; \alpha\, b \bmod q, \qquad a^b \bmod p \;=\; g^\gamma \bmod p. \tag{2}$$

The Chinese Remainder Theorem provides a ring isomorphism $\mathbb{Z}/q\mathbb{Z} \;\cong\; \prod_{i=1}^{k} \mathbb{Z}/m_i\mathbb{Z}$ for the prime-power factorization $q = \prod_i m_i$. Therefore, the product in Equation 2 splits into k per-component products:

$$\gamma \bmod m_i \;=\; \big((\alpha \bmod m_i)\,(b \bmod m_i)\big) \bmod m_i. \tag{3}$$

Each of them involves only the residues of the two operands modulo m_i. We will refer to this process as the *CRT decomposition*, and given a number γ, the combined residues from each m_i as the CRT *tuple*. The model computes the answer by computing each residue $\gamma \bmod m_i$ on its own and reconstructing γ from the resulting tuple.

An important point on prime-power CRT components is that they do not split further; i.e., $9 = 3^2$ remains a single component, the cyclic group $\mathbb{Z}/9\mathbb{Z}$, instead of $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$. What the model does inside such a component is treated separately, in the prime-power section below.

Table 1 instantiates this for the three primes used in the rest of the paper.

Table 1: The CRT components of the three primes analyzed in this paper, and the Fourier band each component occupies. Every entry follows from the factorization of $q = p - 1$; no model weight is involved.

p	$q = p - 1$	component m_i	band $\{\, j\,(q/m_i) \,\}$
83	$82 = 2 \times 41$	2	41
		41	$2, 4, 6, \ldots, 40$
127	$126 = 2 \times 3^2 \times 7$	2	63
		9	$14, 28, 42, 56$
		7	$18, 36, 54$
131	$130 = 2 \times 5 \times 13$	2	65
		5	$26, 52$
		13	$10, 20, 30, 40, 50, 60$

3.2 THE CHARACTER BASIS AND THE PREDICTED FREQUENCIES

The characters of the cyclic group $\mathbb{Z}/q\mathbb{Z}$ are the functions $\chi_k(x) = \exp(2\pi i k x/q)$ for $k = 0, \ldots, q-1$. They form an orthonormal basis for the functions on $\mathbb{Z}/q\mathbb{Z}$. We refer to k as a frequency and to the expansion of a function in this basis as its spectrum.

A function of $\gamma \bmod m$ on $\mathbb{Z}/q\mathbb{Z}$ has period m. Such a function has χ_k in its expansion only when $\exp(2\pi i k m/q) = 1$. That means only when k is a multiple of q/m.

Discarding the constant term and using the conjugate symmetry of a real spectrum, the frequencies of component m are

$$F_m \;=\; \big\{\, j\,(q/m) \;:\; 1 \le j \le \lfloor m/2 \rfloor \,\big\}. \tag{4}$$

We call the union $\bigcup_i F_{m_i}$ the *CRT-predicted frequency set*. The right-hand column of Table 1 is Equation 4 evaluated on the factorization of q. Because the m_i are pairwise coprime, the bands F_{m_i} are disjoint as a frequency cannot belong to two components at once.

Finally, to analyze the model's components in $\mathbb{Z}/q\mathbb{Z}$, the values must be re-indexed following the discrete logarithm mapping $\log_g \colon ((\mathbb{Z}/p\mathbb{Z})^*, \times) \to (\mathbb{Z}/q\mathbb{Z}, +)$, following Nguyen (2026).

4 THE MODEL COMPUTES ON THE CRT BANDS

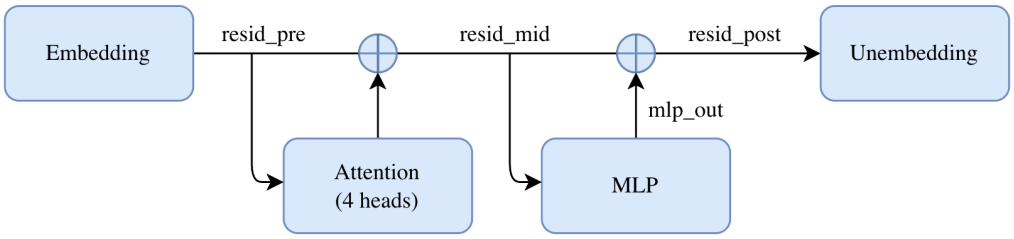


Figure 1: Residual stream positions of interest in the transformer architecture

A linear probe is an affine map $x \mapsto Wx+c$ trained on a frozen activation. A high probe reading at a location implies that the target is linearly accessible there. We train 2 probes per CRT component m along with controls to assess 2 residual stream locations: `resid_mid` and `resid_post` (Figure 1).

Training: For component m, the probe is a single affine layer with m output classes and a bias. It is trained by cross-entropy with Adam for 500 full-batch steps; and the label of an input is the residue itself. The train-test splits for the m_i probes of a model follow their model's original $60/40$ split. Alongside the CRT components we probe a *control modulus*, the smallest prime not dividing q. Since its frequency q/m is not an integer, no character of $\mathbb{Z}/q\mathbb{Z}$ can carry that residue.

Before we can discuss results, it is important to establish what the *baseline chance* is for such a probe.

4.1 WHAT COUNTS AS CHANCE FOR A PER-COMPONENT DECODE

Several of the tests that follow report a *per-component decode*. Writing $\hat{\gamma}$ for the discrete logarithm of the answer the model predicts,

$$A_m \;=\; \Pr\left[\,\hat{\gamma} \equiv \gamma \pmod{m}\,\right], \tag{5}$$

the fraction of inputs whose predicted answer carries the correct residue modulo m. Once a component has been removed the model retains no information about $\gamma \bmod m$. The model's output distribution will be defined by the choice and interaction between the inputs. As shown below, this distribution is not uniform, with the highest probability at class 0, which is the *majority residue* of mod-m stream. This is the *baseline chance* for a CRT stream m, and it is not $1/m$. Appendix B provides a full derivation of the majority residue rate and why it is the baseline chance. Table 2 evaluates it for every CRT component that appears in this paper.

Table 2: The baseline chance of the per-component decode A_m, from Equation 9.

component m	2	4	5	7	8	9	13	16	41
baseline chance	0.750	0.500	0.360	0.265	0.312	0.259	0.148	0.188	0.048

A control modulus c does not divide q, so neither our mechanism framework nor the derivation above applies. There, we report the empirical majority rate, which is close to $1/c$.

4.2 THE PRODUCT IS FORMED IN THE MLP

Using the set of linear probes at `resid_mid`, `resid_post` for each m_i per model (and control), we probe those positions over the entire training process. The result is consistent: the before-MLP readout stays relatively low and flat, moving by less than 0.07 over the entire run; while the after-MLP readout starts at the same level and climbs to 1 as the model groks. Both position control probes' readout stays low, close to baseline chance throughout. Figure 2 and Appendix E show the

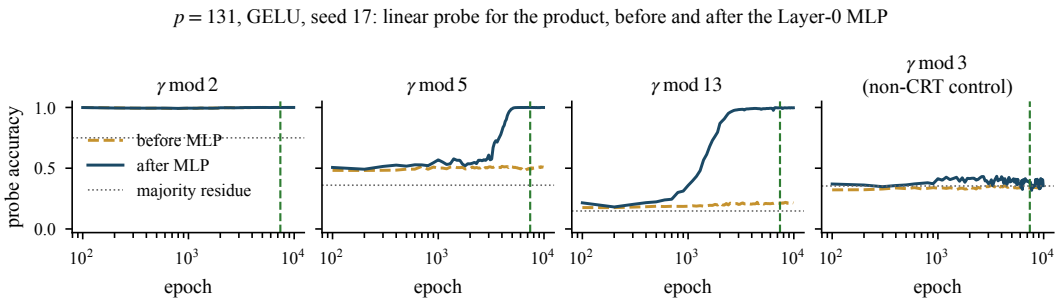


Figure 2: Linear probe for $\gamma \bmod m$ over training, before (gold, dashed) and after (teal, solid) the MLP, read at the answer position. Green line: the epoch at which the model first reaches 100% on all pairs. The dotted line is the majority-residue rate, which is the baseline chance for each component respectively.

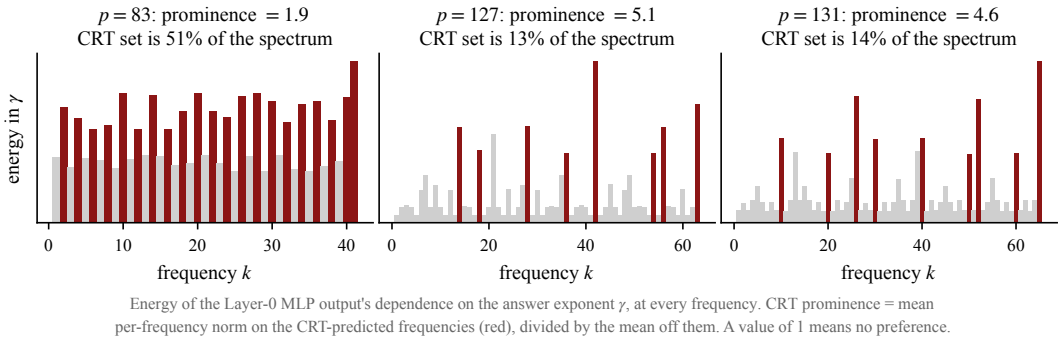


Figure 3: Energy of the MLP output's dependence on the answer exponent γ, at every frequency. Red bars are the CRT-predicted frequencies, named from the factorization of q alone.

probe readout evolution throughout the training process. The exceptions are the mod-2 streams. All mod-2 probes' readouts are near 1 for both residual stream positions, throughout the training process, and across all models. We dedicate Appendix F to go in-depth on this. Nevertheless, the following sections will use causal methods to show that the mod-2 stream is calculated in the MLP.

These results do not imply that the model produces these residues through a linear process. What they show is that the CRT streams' results are all perfectly linearly accessible at `resid_post`. And this is consistent with the claim that W_U combines the residues to produce the final answer.

4.3 The answer lives on the predicted bands

To identify the frequencies that represent the MLP answer, we take the MLP output, mean-center it over the dataset, and measure its dependence on γ at every frequency $k = 1, \ldots, \lfloor q/2 \rfloor$. This is done by projecting it onto $\cos(2\pi k\gamma/q)$ and $\sin(2\pi k\gamma/q)$ and combining the two norms. We summarize the result as the *CRT prominence*: the mean of the per-frequency norms on the CRT-predicted set divided by the mean of the norm of the non-CRT set. A value of one means the predicted frequencies carry no more energy than the rest.

Figure 3 shows the spectra. The prominence is 1.9 to 2.0 for $p = 83$, 4.4 to 5.1 for $p = 127$ and 4.5 to 4.7 for $p = 131$. Each model gives its own spectrum and the range spans the three seeds of that prime; the per-model values are in Table 3. They are consistently greater than 1.

4.4 So are the weights

Three linear maps in the forward pass are indexed by a value token and can therefore be read spectrally: **1.** The embedding W_E has one row per token. **2.** The unembedding W_U has one column per

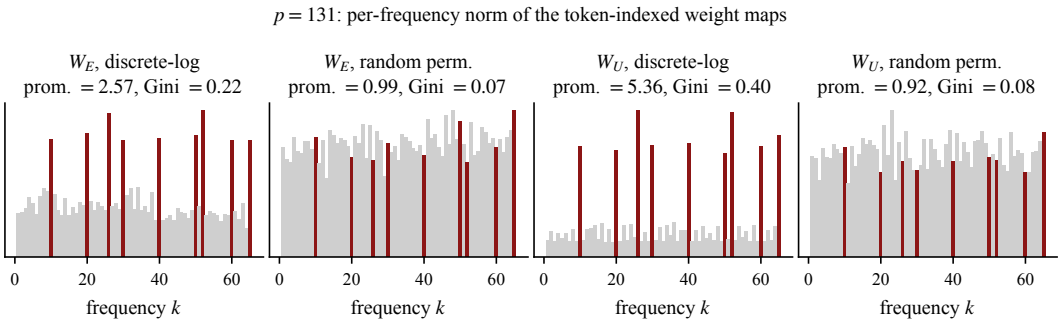


Figure 4: Per-frequency norm of W_E and W_U with rows in discrete-log order (top) and in one random token permutation. The permutation is the control: the CRT-predicted frequencies are not prominent in an arbitrary ordering. Other primes in Appendix D.

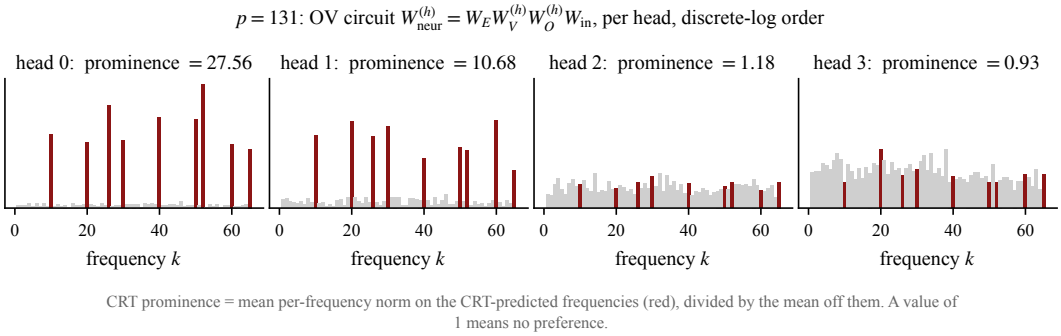


Figure 5: The four OV circuits read separately. Two heads carry the CRT structure and two do not, a pattern that holds in all nine models.

candidate answer, transposed so its rows are indexed by tokens. **3.** Finally, following Nanda et al. (2023), the composed read-in path of head h is $W^{(h)}_{\text{neur}} = W_E W^{(h)}_V W^{(h)}_O W_{\text{in}}$, which maps a token to the MLP pre-activations and so carries whatever that head delivers into the MLP. For each map we place the rows in discrete-log order (Nguyen, 2026), project onto the character basis, and take the combined per-frequency norm.

Extending the idea from Nguyen (2026), we are showing that not only are the spectra sparse under logarithm-reindexing, but also that the "key frequencies" are exactly within the CRT-predicted set. This shows that the mechanism is working in the *multiplicative basis*, and the irreducible representations' frequencies are the CRT-predicted ones. Logarithm-reindexing is one permutation of the tokens. For visual comparison we offer a single random permutation as control (Figure 4); and, to make it rigorous, we use a *null distribution* over 200 random orders (Figure 18). The discrete-log order places the CRT prominence of W_E and W_U between 14 and 85 null standard deviations above the null mean, above every one of the 200 draws, an empirical $p \leq 0.005$ in all nine models. The in-depth statistical results can be found in Appendix I. The reindexing also concentrates the spectrum as both CRT prominence and Gini coefficient rises (Table 3).

Figure 5 plots the four OV circuits separately. We call a head CRT-aligned when its prominence exceeds all 200 null draws for that head. By that test, exactly *two of the four heads are aligned* in every one of the nine models. Statistics results for the heads can also be found in Appendix I. We conclude that the CRT structure in the read-in path is carried by two heads per model while the other two are, by this measure, indistinguishable from an arbitrary token order. Which two they are is seed-dependent.

4.5 CAUSAL INTERVENTIONS

Because each CRT band carries an arithmetic label, we can predict the consequence of causal intervention.

The intervention site: The following tests act on the MLP output at the answer position, hooking `blocks.0.mlp.hook_post`. For a component m we build an orthonormal basis U of the span of $\{\cos(2\pi k\gamma/q), \sin(2\pi k\gamma/q) : k \in F_m\}$ over the dataset, with the columns mean-centered. Because the model has no LayerNorm, removing a subspace from the MLP output leaves the skip path untouched. This is the intervention Nanda et al. (2023) use for excluded loss, applied on each CRT component.

The measurement: A_m is the fraction of inputs for which $\log_g$ of the predicted answer is congruent to γ modulo m (Equation 5). Since all models grok, before any intervention, A_m is 1 for every CRT stream. The *baseline chance* for a model to produce a correct prediction $\gamma \bmod m$ follows Table 2 (with derivation in Appendix B).

Necessity: Ablate F_m and decode every component. Figure 6 gives the matrices. Removing a component's band causes the component's decode to drop to its baseline chance while leaving the others' virtually untouched, including the controls. Over all nine models the smallest off-diagonal entry is 0.960, and the non-CRT control row never falls below 0.998. The mod-2 entry is between 0.749 and 0.763 across the nine models, against a majority rate of 0.750. This applies to almost every other component also. Five components depart from it, all discussed in Appendix J.

Sufficiency: The following test rules out the possibility of a band being an incidental carrier. For each input we replace the content of band F_m in the MLP output with that of a *CRT-counterfactual* source. A counterfactual source γ_s of target γ_t shares the same CRT decomposition tuple except at one position. For example, in $p = 131$, the CRT tuple is $(\gamma \bmod 2, \gamma \bmod 5, \gamma \bmod 13)$; the target can be $0 \leftrightarrow (0, 0, 0)$ and the source can be $26 \leftrightarrow (0, 1, 0)$. If the band carries component m of the answer, the prediction's residue mod m should follow the source while every other residue stays with the target. Figure 6 shows this is what happens. The transplanted residue follows the source on 0.762 to 1.000 of inputs across the nine models. Other component keeps the target at 0.953 or above, and the non-CRT control rows at 0.968 or above, with one exception reported in Appendix J.

Restriction: Ablation and patching both leave the rest of the MLP output in place. The *complementary test* keeps only a band and retains each neuron's means while discarding the orthogonal complement of the MLP output. If the read-out reconstructs the answer from the bands, then restricting to the CRT bands should preserve the answer. More fine-grained, restricting to a CRT band preserves only that component's decode. Figure 17 shows that **both** hold: keeping all CRT bands preserves the answer at 0.966 to 1.000 joint accuracy across the nine models, against 0.003 to 0.018 for the same restriction on a randomly initialized network. Appendix H gives the matrices and in-depth discussion on the method.

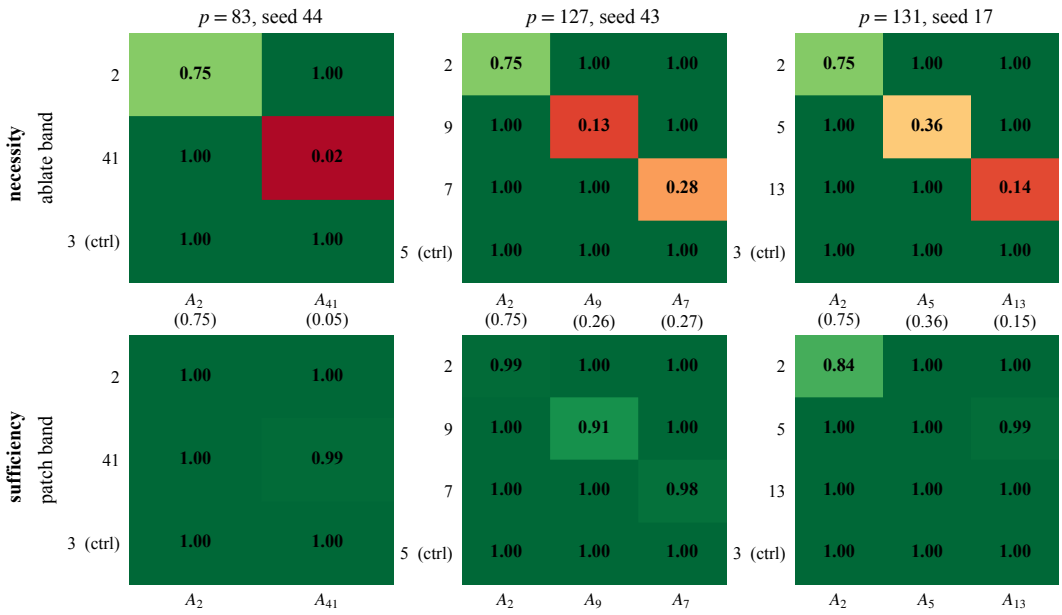
These tests show that in a CRT stream m, the band carries the answer's residue mod m, and that the CRT bands together carry the whole answer.

5 INSIDE A PRIME POWER: A QUOTIENT FILTRATION

The operation inside a CRT stream is multiplication; therefore, the Clock mechanism (Zhong et al., 2023) does not apply directly. For prime m streams, a second logarithm is required to turn the product into a sum $L_\gamma = L_\alpha + L_b \bmod \phi(m)$ (Nguyen, 2026). Even then, analysis shows that the behavior is not always a "clean" Clock, but rather a mixture of Clock and Pizza algorithms. Appendix G gives the diagnostics per component. For prime power components, (i.e., $m = 9, 16$), the second log trick above does not work or even applicable at all. Those components carry a graded structure of their own, which is the novel finding of this section.

Consider a prime power CRT component $m = \ell^e$, we want to show that its irreducible representations (*irreps*, or simply *characters*) form a filtration that is *dual* to:

$$\mathbb{Z}/\ell^e\mathbb{Z} \twoheadrightarrow \mathbb{Z}/\ell^{e-1}\mathbb{Z} \twoheadrightarrow \cdots \twoheadrightarrow \mathbb{Z}/\ell\mathbb{Z}, \tag{6}$$



Top: ablate band F_m out of the MLP output, then decode each component; the baseline chance is in brackets. Bottom: overwrite F_m with that of a source input differing only mod m; the diagonal is the fraction that followed the SOURCE, off-diagonal the fraction that kept the TARGET.

Figure 6: Necessity (top) and sufficiency (bottom) on the same axes. Ablating a band destroys exactly the residue it is predicted to carry; patching it from a CRT-counterfactual source transplants exactly that residue and leaves every other component with the target. All nine models are in Appendix K.

Equivalently, every irreducible representation of a smaller quotient is also one of the larger group. Now we know that the characters of a CRT component m are indexed by its CRT frequencies. The *order* of a character with frequency j (q/m) is $d = m/\gcd(j, m)$.

Take $p = 113$, where $q = 112 = 2^4 \times 7$ and the CRT component $m = 16$ has the widest band in this paper. Its eight predicted frequencies j $(q/16) = 7j$ split by order is:

$$\underbrace{\{56\}}_{\text{order 2}}, \quad \underbrace{\{28\}}_{\text{order 4}}, \quad \underbrace{\{14, 42\}}_{\text{order 8}}, \quad \underbrace{\{7, 21, 35, 49\}}_{\text{order 16}}. \tag{7}$$

To show the *duality* above, we will ablate each order's frequencies. Figure 7 gives the result, and it is a staircase. That means it supports our claim: Destroying an order affects its own and higher order's decode while leaving the lower order's virtually untouched. The other supporting evidence is that the lower order ablations affect more decodes, but each to a lesser degree. This is because the spans of order 2, 4, 8, 16 are 1, 2, 4, 8, respectively. A more technical mathematical discussion and other prime power components ablation results can be found in Appendix L.

6 Training Dynamics: Carving Out the Circuit

Inspired by the "restricted loss" and "excluded loss" progress measures of Nanda et al. (2023), we provide our own way to observe the circuit formation. Let the *generalizing circuit* be the aggregated set of weights that the information flows through once the model has generalized, that is $W_E \to \text{OV} \to \text{MLP} \to W_U$. As shown above, the CRT frequencies within a model are labeled frequencies across structures from the factorization of q alone. Section 4.5 shows causally that the answer is computed on these frequencies. If they are the generalizing circuit, then they should also be what weight decay preserves through the *cleanup* phase of grokking.

Let the *normalized-weight-norm* of the CRT circuit be defined as follows. For each linear map of Section 4.4 we place its rows in discrete-log order, take the Fourier transform along the token

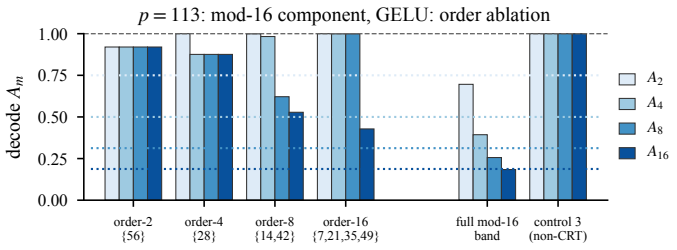


Figure 7: Ablating order 2, 4, 8, 16 brings their decode accuracy down to 0.920, 0.876, 0.622, 0.428 respectively while leaving the lower order's at 0.98 or above. Other models' reports are in Appendix L.

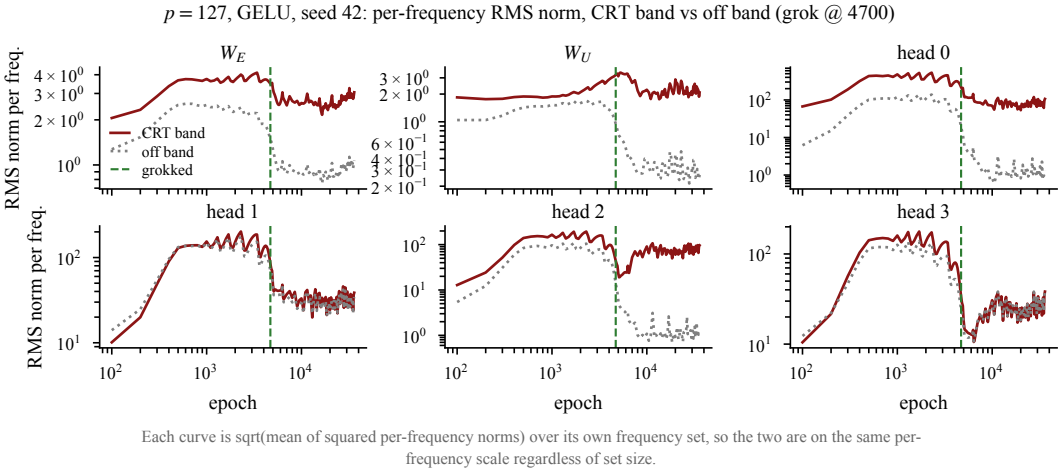


Figure 8: Weight-norms of CRT band vs off-band of model $p = 127$. CRT band's consistently sit higher than off-band's in the structures that carry this circuit. Heads 1 and 3 hover at parity throughout as they are the two this model does not use. Other models are in Appendix M.

axis, and record the norm of each frequency component. The normalized-weight-norm of a set of frequencies is the root-mean-square of those per-frequency norms. Here, we are interested in the CRT frequency set (or *on-band*) and its complement (or *off-band*). Figure 8 shows how the two weight-norms evolve during the training process. Under weight decay, both norms tend to shrink. However, like carving stones, the on-band norms always come out ahead.

7 Discussion and Conclusion

Why the bands are canonical. For modular addition (Nanda et al., 2023), every character is a homomorphism $\chi_k(\alpha + b) = \chi_k(\alpha)\,\chi_k(b)$. So any frequency that is coprime to p can carry the whole answer. A product has no such identity. Therefore, the CRT bands acting as key frequencies is a phenomenon that is consistent across models and seeds.

Limitations. We do not analyze the attention circuit, and how W_E comes to be ordered by the discrete logarithm remains open. Inside a prime power we identify its irreducible representations, but have not yet found the algorithm that computes them. Two of the three trajectories in Section 6 are GELU while the analysis set is ReLU. Whether such a mechanism arises in larger models is unknown.

This paper reverse engineers how one-layer transformers compute $a^b \bmod p$. The bands follow from factoring q and each carries an arithmetic label. Therefore, every causal intervention and the generalizing circuit formation can be predicted before any weight is inspected.

AI USE STATEMENT

Generative AI is used extensively throughout this project. It is used for research ideation and execution; retrieval and discovery of related works; aid in producing the first draft of this paper and polish writing. The authors ideate the original idea and direct the research; check the code quality and its result; read and confirm references' correctness. The authors confirm that there is no plagiarism as all research directions are human-ideated, the used methods are public and well-established, the borrowed ideas are referenced. Most importantly, the final draft of the paper is human-written by the authors as all steps and experiments are vetted and checked again. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The entire project code can be found anonymously at `https://osf.io/fcw4a/?view_only=9c3df3f0a3184a9093d802bca1f7d5b8`, together with the model weights and the cached intermediate results. Every figure and table in this paper regenerates from those cached results with no model download. The code responsible for training the models is also provided. However, the original models are trained on Apple's MPS, which is notable for its indeterministic behavior. Given the exact same seed, a newly trained model may not be exactly the same as the one used in this paper.

REFERENCES

David Demitri Africa, Sara M. Kapoor, Theo Simon Sorg, and Challenger Mishra. Learning modular exponentiation with transformers. *arXiv preprint arXiv:2506.23679*, 2025.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning*, 2023.

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Towards empirical interpretation of internal circuits and properties in grokked transformers on modular polynomials. *Transactions on Machine Learning Research*, 2024.

Gavin McCracken, Gabriela Moisescu-Pareja, Vincent Letourneau, Doina Precup, and Jonathan Love. Uncovering a universal abstract algorithm for modular addition in neural networks. *arXiv preprint arXiv:2505.18266*, 2025.

Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Huu Danh Nguyen. The discrete-log clock: How a transformer learns modular multiplication. In *ICML 2026 Workshop on Mechanistic Interpretability*, 2026.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR 2022 Workshop on MATH-AI*, 2022.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *International Conference on Machine Learning*, 2024.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, 2023.

A PRELIMINARIES AND SETUP

Task. We fix a prime modulus p and train on the full table of $a^b \bmod p$ for $a, b \in \{1, \ldots, p-1\}$, which is $(p-1)^2$ input pairs. The value 0 is excluded: it is not in the multiplicative group and has no discrete logarithm. A single example is the token sequence $[a, b, =]$ with a dedicated token for $=$, and the answer is read from the logits at the third position over the $p-1$ candidate values.

Model. All models analyzed here are one-layer decoder-only transformers in the style of Nanda et al. (2023), implemented with TransformerLens: 4 attention heads, $d_{\text{model}} = 512$, $d_{\text{head}} = 128$, $d_{\text{mlp}} = 512$, no LayerNorm, and all biases frozen at initialization. Removing LayerNorm keeps the residual stream a plain sum of the attention output, the MLP output and the skip, which is what lets an intervention on the MLP output be read as an intervention on one additive term.

Training. AdamW with learning rate 10^{-3}, weight decay 1.0, $\beta = (0.9, 0.98)$, full-batch, on a random 60% of the pairs. We call a model grokked when it reaches 100% accuracy on all $(p-1)^2$ pairs, not on the held-out split alone, and sustains it over five consecutive checks. Twenty one-layer models reached that criterion across moduli and seeds.

The analysis set. The nine models used throughout are ReLU models at three primes with three seeds each: $p = 83$ (seeds 44, 45, 46), $p = 127$ (43, 45, 46) and $p = 131$ (17, 67, 69). These three primes were chosen for the arithmetic of $q = p - 1$ rather than for their behaviour: $82 = 2 \times 41$ is the minimal two-component case, $130 = 2 \times 5 \times 13$ is square-free with three components, and $126 = 2 \times 3^2 \times 7$ carries the prime power. A fourth prime, $p = 113$ with $q = 112 = 2^4 \times 7$, is used only for the four-level power-of-two filtration, in both a ReLU and a GELU model. Every model reported is at 100.00% on all pairs at the checkpoint analyzed.

Snapshots. Training snapshots were kept for one trajectory per prime: $p = 83$ ReLU seed 44 (169 snapshots), $p = 127$ GELU seed 42 (124) and $p = 131$ GELU seed 17 (99). Only the $p = 83$ trajectory shares an activation function with the analysis set; the dynamics section states this where it matters.

Determinism. Training and analysis ran on Apple MPS, whose reductions are not bitwise deterministic, so exact replication of a given seed's trajectory is not guaranteed. The qualitative results reported here reproduce across seeds and primes.

B CRT COMPONENT DECODE BASELINE CHANCE

For a sum the residue would be uniform: if α is uniform on $\mathbb{Z}/m\mathbb{Z}$ then so is $\alpha + b$ for any independent b, which is why $1/m$ is the correct reference on modular addition. A product is not uniform, because $\gamma \equiv 0$ whenever *either* factor is divisible by m. The clearest case is $m = 2$, where αb is odd only if α and b are both odd. A model that has lost the mechanism to compute $\gamma \bmod 2$ gives the answer 0 $3/4$ of the time and be correct. This is because 0 has a chance of $3/4$ and is the mode of this distribution and is the *majority residue* of $\gamma \bmod 2$.

Take $p = 83$ as an example. Exactly 41 of the 82 values of α are odd, and likewise for b, so

$$\Pr[\,\gamma \text{ odd}\,] \;=\; \frac{41 \times 41}{82^2} \;=\; \frac{1681}{6724} \;=\; \frac{1}{4}, \qquad \Pr[\,\gamma \text{ even}\,] \;=\; \frac{3}{4}. \tag{8}$$

The general statement follows from the ℓ-adic valuation v_ℓ: Every CRT component is a prime power $m = \ell^e$ dividing q, and $\log_g$ is a bijection onto $\mathbb{Z}/q\mathbb{Z}$, while b runs over a complete residue system modulo q. Therefore, $\alpha \bmod m$ and $b \bmod m$ are independent and uniform on $\mathbb{Z}/m\mathbb{Z}$. Their product vanishes exactly when the two valuations sum to at least e. Counting the pairs by valuation gives the whole distribution: for a residue $r \neq 0$ with $j = v_\ell(r) < e$,

$$\Pr[\,\gamma \equiv 0\,] \;=\; \frac{\ell + e\,(\ell - 1)}{\ell^{\,e+1}}, \qquad \Pr[\,\gamma \equiv r\,] \;=\; \frac{(j+1)(\ell-1)}{\ell^{\,e+1}}. \tag{9}$$

The first expression always exceeds the second, so the modal residue is $r = 0$ and the baseline chance of A_m is the left-hand quantity. Setting $e = 1$ recovers $(2m-1)/m^2$ for a prime component, and $\ell = 2$, $e = 1$ recovers the three quarters of Equation 8.

To summarize, the distribution of the output $\gamma \bmod m$ is not uniform, with the majority rate at class 0 is because the choices we made on the inputs: α uniformly sampled, b is independent, and their interaction is modular multiplication.

For a linear probe, its output distribution will converge to the model's true label prior due to training with cross-entropy. And, with argmax, the probe's baseline chance will match the *majority residue* of mod-m stream.

For tests using ablation, by remove the mechanism that the model uses to compute mod-m stream, the model's output distribution (logits) also converges to the true label prior. And, with argmax, the ablated model's baseline chance will match the *majority residue* of mod-m stream.

C ALL NINE MODELS

Table 3 gives the CRT prominence of every token-indexed map for all nine grokked ReLU models, under the discrete-log order and under one random token permutation. Table 4 gives the full necessity and sufficiency results per component. Table 5 gives the clock diagnostics, and Table 6 the per-level filtration ablations including the GELU model at $p = 113$.

Table 3: All nine grokked ReLU models. Columns 2 to 5 are the CRT prominence, the mean per-frequency norm on the CRT-predicted frequencies divided by the mean off them. Columns 6 to 9 are the Gini coefficient of the same spectrum, which measures how concentrated it is overall rather than where the energy sits. "log" places the rows in discrete-log order, "perm" in one random token permutation. "best head" is the largest per-head OV prominence, and the last column is the prominence of the Layer-0 MLP output's dependence on γ. Note that the Gini values are far below what a spectrum supported only on the CRT set would give (0.49 for $p = 83$, 0.87 for $p = 127$, 0.86 for $p = 131$): the energy is preferentially on the predicted frequencies, but the spectra are not sparse in absolute terms.

	CRT prominence				Gini					
model	W_E	perm	W_U	perm	W_E	perm	W_U	perm	best head	MLP out
p83 seed 44	1.62	0.98	4.00	1.02	0.173	0.073	0.322	0.083	35.19	1.91
p83 seed 45	1.59	0.97	1.95	1.00	0.152	0.082	0.180	0.061	70.21	1.99
p83 seed 46	1.71	0.94	2.99	0.93	0.176	0.087	0.266	0.084	28.41	2.03
p127 seed 43	2.37	1.02	3.10	1.01	0.173	0.050	0.239	0.045	13.23	5.06
p127 seed 45	2.52	1.05	5.40	0.99	0.207	0.060	0.386	0.068	7.97	4.64
p127 seed 46	2.53	0.99	4.00	1.02	0.195	0.053	0.338	0.055	16.53	4.39
p131 seed 17	2.57	0.99	5.36	0.92	0.219	0.067	0.404	0.078	27.56	4.62
p131 seed 67	2.27	1.05	3.89	1.00	0.188	0.058	0.397	0.086	10.04	4.71
p131 seed 69	2.12	1.02	3.41	1.02	0.169	0.050	0.304	0.064	9.97	4.45

Table 4: Necessity and sufficiency, per component, for all nine models. "ablate F_m" is A_m after its own band is removed; "chance" is the majority-residue rate of $\gamma \bmod m$, not $1/m$. "worst off-diag." is the smallest A_m observed when some OTHER band is removed (higher is better). "patch F_m" is the fraction of inputs whose residue mod m followed the counterfactual source. "control row" is A_m when a non-CRT band is removed.

model	m	chance	ablate F_m	worst off-diag.	patch F_m	control row
p83 seed 44	2	0.750	0.749	1.000	1.000	1.000
	41	0.048	0.022	1.000	0.992	1.000
p83 seed 45	2	0.750	0.750	1.000	1.000	1.000
	41	0.048	0.048	1.000	0.997	1.000
p83 seed 46	2	0.750	0.751	0.960	0.998	1.000
	41	0.048	0.047	1.000	0.969	1.000
p127 seed 43	2	0.750	0.750	1.000	0.991	1.000
	9	0.259	0.132	0.997	0.912	1.000
	7	0.265	0.277	1.000	0.979	1.000
p127 seed 45	2	0.750	0.750	0.998	0.991	1.000
	9	0.259	0.080	0.985	0.942	1.000
	7	0.265	0.190	1.000	0.950	1.000
p127 seed 46	2	0.750	0.763	1.000	0.988	1.000
	9	0.259	0.156	0.971	0.867	0.998
	7	0.265	0.286	0.996	0.925	1.000
p131 seed 17	2	0.750	0.750	1.000	0.844	1.000
	5	0.360	0.357	0.997	0.999	1.000
	13	0.148	0.144	1.000	1.000	1.000
p131 seed 67	2	0.750	0.750	1.000	1.000	1.000
	5	0.360	0.840	0.980	0.762	1.000
	13	0.148	0.170	1.000	0.997	1.000
p131 seed 69	2	0.750	0.750	1.000	1.000	1.000
	5	0.360	0.546	1.000	0.946	1.000
	13	0.148	0.231	1.000	0.988	1.000

Table 5: Clock diagnostics per cyclic component. ρ is the share of the top band neuron's variance explained by $\gamma \bmod m$; top-1 is the largest harmonic's share of the residue-pooled response over the units. "n." is the tuned-neuron grid and "lg." the correct-logit grid, with f_{s} and f_{d} the sum-diagonal and difference-antidiagonal energy shares. Entries marked "undef." are cases where the statistic is not defined: the mod-9 neuron is identically zero on the units, and q is uninformative when the dominant sum harmonic j^* sits at the Nyquist bin $\varphi(m)/2$.

p	m	φ	ρ	top-1	n. f_{s}	n. f_{d}	lg. f_{s}	lg. f_{d}	q	j^*
p = 83	41	40	0.14	0.95	0.11	0.11	0.97	0.00	1.00	14
p = 127	7	6	0.20	1.00	0.14	0.03	0.17	0.58	0.59	2
p = 127	9	6	0.79	undef.	undef.	undef.	0.46	0.04	undef.	3
p = 131	5	4	0.60	0.63	0.03	0.03	0.70	0.12	undef.	2
p = 131	13	12	0.20	0.63	0.09	0.08	0.91	0.04	undef.	6

Table 6: Prime-power filtration: per-order ablation. Columns are the decoded moduli in the order given in the text ($3, 9$ for p127; $2, 4, 8, 16$ for p113). "rank" is the number of real directions the ablated band spans over the batch.

model	ablated order	rank	A col 1	A col 2	A col 3	A col 4
p127 s43	order-3 42	2	0.773	0.773		
	order-9 14,28,56	6	1.000	0.157		
	full mod-9 band	8	0.559	0.132		
	control 5 (non-CRT)	4	1.000	1.000		
p127 s45	order-3 42	2	0.778	0.778		
	order-9 14,28,56	6	1.000	0.255		
	full mod-9 band	8	0.580	0.080		
	control 5 (non-CRT)	4	1.000	1.000		
p127 s46	order-3 42	2	0.718	0.718		
	order-9 14,28,56	6	0.938	0.308		
	full mod-9 band	8	0.502	0.156		
	control 5 (non-CRT)	4	1.000	0.998		
p113 relu	order-2 56	1	1.000	1.000	1.000	0.999
	order-4 28	2	0.997	0.745	0.741	0.741
	order-8 14,42	4	1.000	1.000	0.633	0.583
	order-16 7,21,35,49	8	1.000	1.000	0.998	0.442
	full mod-16 band	15	0.842	0.440	0.222	0.165
	control 3 (non-CRT)	2	1.000	1.000	1.000	1.000
p113 gelu	order-2 56	1	0.920	0.920	0.920	0.920
	order-4 28	2	1.000	0.876	0.876	0.876
	order-8 14,42	4	0.999	0.983	0.622	0.528
	order-16 7,21,35,49	8	1.000	0.999	0.999	0.428
	full mod-16 band	15	0.696	0.393	0.256	0.184
	control 3 (non-CRT)	2	1.000	1.000	1.000	1.000

D WEIGHT SPECTRA AT THE OTHER PRIMES

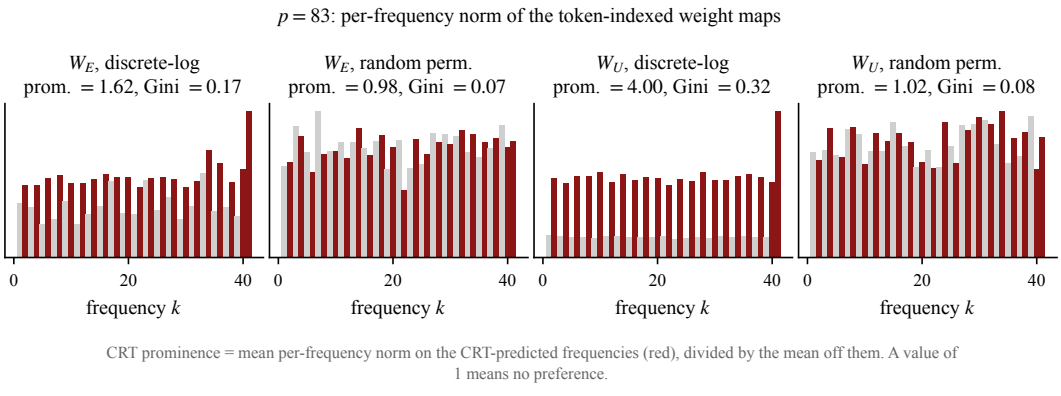


Figure 9: $p = 83$: per-frequency norm of W_E and W_U, discrete-log order against one random token permutation. Note that for this prime the CRT-predicted set is 21 of 41 frequencies, which caps how large a prominence ratio can be.

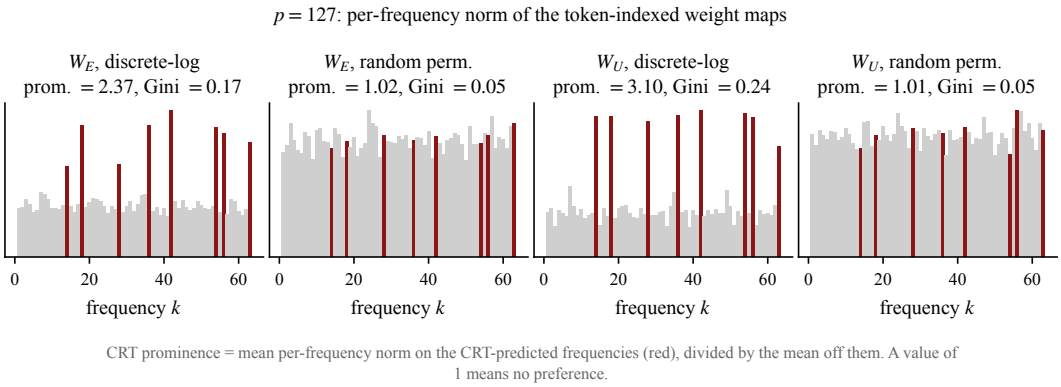


Figure 10: $p = 127$: per-frequency norm of W_E and W_U, discrete-log order against one random token permutation.

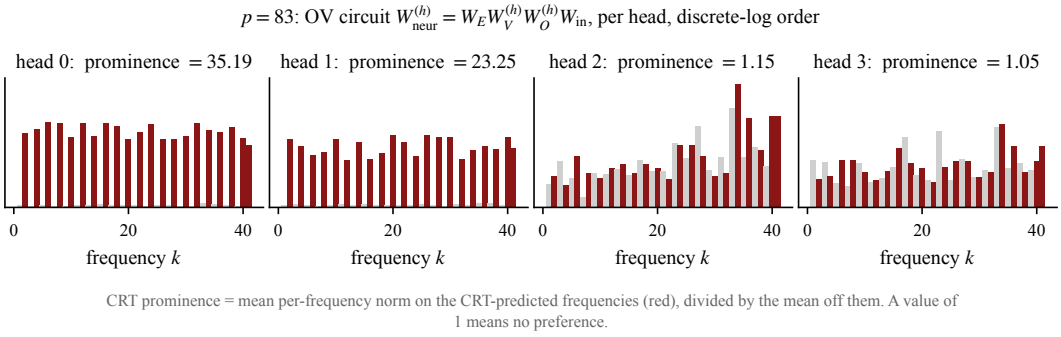


Figure 11: $p = 83$: the four OV circuits read separately.

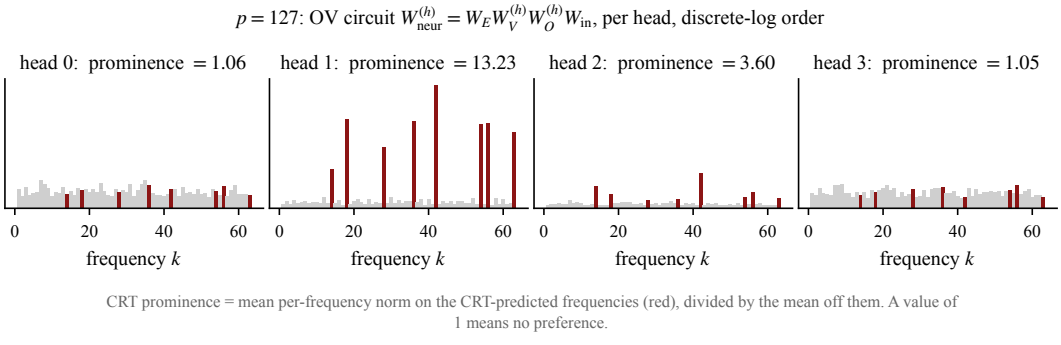


Figure 12: $p = 127$: the four OV circuits read separately.

E GROKKING CURVES AND THE OTHER PROBE TRAJECTORIES

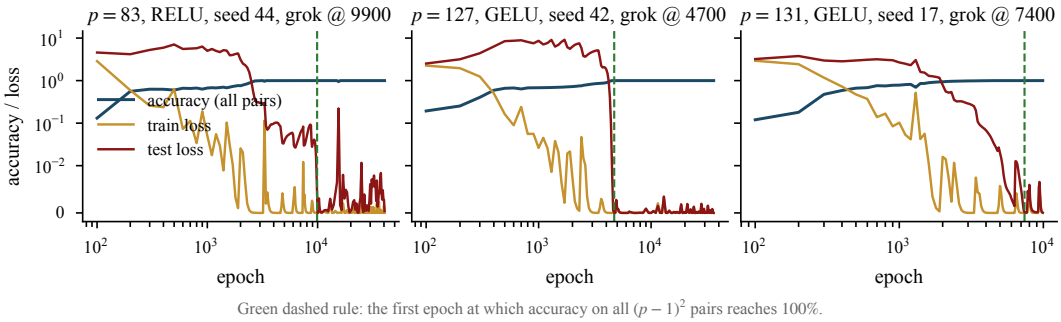


Figure 13: The three snapshot trajectories. Only the $p = 83$ model uses ReLU.

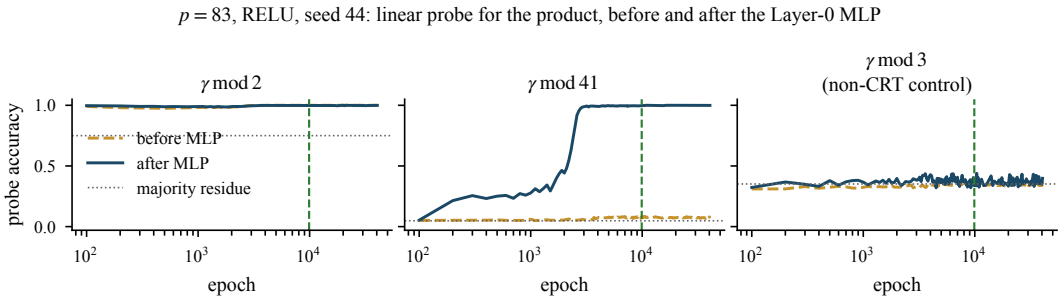


Figure 14: $p = 83$ ReLU: probe for $\gamma \bmod m$ before and after the Layer-0 MLP over training.

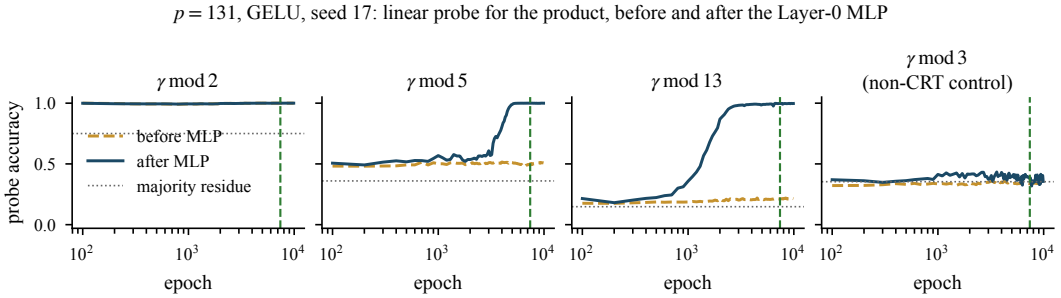


Figure 15: $p = 131$ GELU: probe for $\gamma \bmod m$ before and after the Layer-0 MLP over training.

F THE MOD-2 STREAM IS READABLE BEFORE THE MLP

The mod-2 stream's probe reads near 1 before the MLP as well as after, throughout training and in every model. Across the three trajectories the pre-MLP reading stays in $[0.975, 1.000]$ for $p = 83$, $[0.996, 1.000]$ for $p = 127$ and $[0.992, 1.000]$ for $p = 131$, against a baseline chance of 0.750 (Figure 2, Appendix E). This is expected, and it is a property of the probe rather than of the model.

Since γ is odd exactly when α and b are both odd, $\gamma \bmod 2$ is the logical AND of the two operand parities. AND is linearly separable while XOR is not (Minsky & Papert, 1969). Both parities are already linearly readable before the MLP (0.99 or above for both $\alpha \bmod 2$ and $b \bmod 2$ on $p = 83$). Therefore, a linear probe recovers any separable function of them with no multiplication: Table 7 gives the full set.

The model nonetheless computes its mod-2 residue in the MLP. Ablating the mod-2 band of the MLP output drops A_2 to between 0.749 and 0.762 across the nine models, against a baseline chance of 0.750. The early reading is therefore linearly accessible but unused, which is exactly the probe false positive that the causal tests in the following sections are there to catch. We do not characterize how attention places the two parity bits in the stream, which would require analyzing the query-key circuit. We do not extend the account to the larger streams, which also sit above their baseline chance before the MLP.

Table 7: Linear probe accuracy for every Boolean function of the two operand parities $\alpha \bmod 2$ and $b \bmod 2$, read before and after the MLP at the answer position, for the display model of each prime. "sep." marks whether the target is linearly separable in those two bits. Every separable target is readable before the MLP; the two that are not separable sit at their baseline until the MLP supplies the interaction.

		$p = 83$		$p = 127$		$p = 131$	
target	sep.	pre	post	pre	post	pre	post
$\alpha \bmod 2$	yes	0.994	0.825	1.000	1.000	1.000	0.754
$b \bmod 2$	yes	1.000	0.753	1.000	1.000	1.000	0.749
$\alpha \wedge b \ (= \gamma \bmod 2)$	yes	1.000	0.999	1.000	1.000	1.000	1.000
$\alpha \vee b$	yes	0.992	0.995	1.000	0.998	1.000	1.000
$(\neg\alpha) \wedge b$	yes	0.987	0.742	1.000	1.000	1.000	0.801
$\alpha \oplus b$	no	0.496	0.970	0.503	0.997	0.484	1.000
$\neg(\alpha \oplus b)$	no	0.492	0.969	0.499	0.997	0.493	1.000

Baseline chance is 0.500 for the four two-class targets that are balanced and 0.750 for the three whose majority residue is $3/4$. The separable and non-separable rows separate cleanly in all three models: every separable target reads between 0.987 and 1.000 before the MLP, while $\alpha \oplus b$ and its negation read between 0.484 and 0.513 against a baseline of 0.500, and only reach 0.969 to 1.000 after the MLP. The ranges quoted here are over repeated runs, because a probe whose target is not present reads its baseline plus noise: the probes sitting at a baseline move by up to 0.06 between runs on non-deterministic hardware, while every reading a claim rests on is at 0.98 or above.

G Clock and Pizza diagnostics per component

Inside a cyclic component of odd prime-power modulus m the unit group $(\mathbb{Z}/m\mathbb{Z})^*$ has a primitive root ψ, so on the units we may take a second logarithm $L_x = \log_\psi(x \bmod m)$ and the per-component product becomes an addition, $L_\gamma = L_\alpha + L_b \bmod \varphi(m)$. Only in that coordinate is a Clock defined. We pool a signal onto the $\varphi(m) \times \varphi(m)$ grid indexed by (L_α, L_b) and split its two-dimensional spectral energy into the share on the sum diagonal (f_{sum}, a function of $L_\alpha + L_b$, the Clock form), the share on the difference antidiagonal (f_{diff}, the Pizza form of Zhong et al. (2023)), and the remainder f_{other}. The three shares sum to one, so reporting a mixture is what the measurement returns rather than a failure to decide.

Two views are needed and they answer different questions. A neuron implementing the trigonometric identity $\cos L_\alpha \cos L_b = \frac{1}{2}[\cos(L_\alpha + L_b) + \cos(L_\alpha - L_b)]$ must carry *equal* sum and difference energy, so a low f_{sum} at the neuron level is the building block working, not a failed clock. The difference terms cancel only once the neurons are recombined through W_{out} and W_U, so the verdict is read at the logit of the correct answer. We omit the gradient-symmetricity test of Zhong et al. (2023): it compares gradients with respect to the two operand embeddings and is informative only when the target is symmetric under exchanging them, which ours is not, since a enters through $\log_g$ and b directly.

Table 5 gives the per-component numbers and Figure 16 the grids. Read at the correct logit, the two large components are clean Clocks: mod-41 of $p = 83$ places 0.97 of its energy on the sum diagonal with 0.03 unexplained, and mod-13 of $p = 131$ places 0.91 with 0.05 unexplained. The smaller ones are genuinely mixed: mod-7 of $p = 127$ leans Pizza ($f_{\mathrm{diff}} = 0.58$ against $f_{\mathrm{sum}} = 0.17$), and mod-5 of $p = 131$ has $\varphi(5) = 4$, which is too small for the difference statistic to be meaningful. The mod-9 component is a separate matter: its top band neuron is exactly zero on all six units and

fires only on $\{0, 3, 6\}$, so the grid is constant and every statistic computed on it is undefined rather than zero. That is the observation the filtration section starts from.

One caveat applies to the distance-irrelevance statistic q. When the dominant sum harmonic sits at the Nyquist bin $j^* = \varphi(m)/2$, $\cos(\pi(L_\alpha + L_b)) = \cos(\pi(L_\alpha - L_b))$ identically, so a pure sum harmonic is indistinguishable from a pure difference harmonic and q carries no information. This is the case for mod-13 of $p = 131$, where $j^* = 6$; we report q as undefined there rather than as evidence against the Clock that $f_{\text{sum}} = 0.91$ indicates.

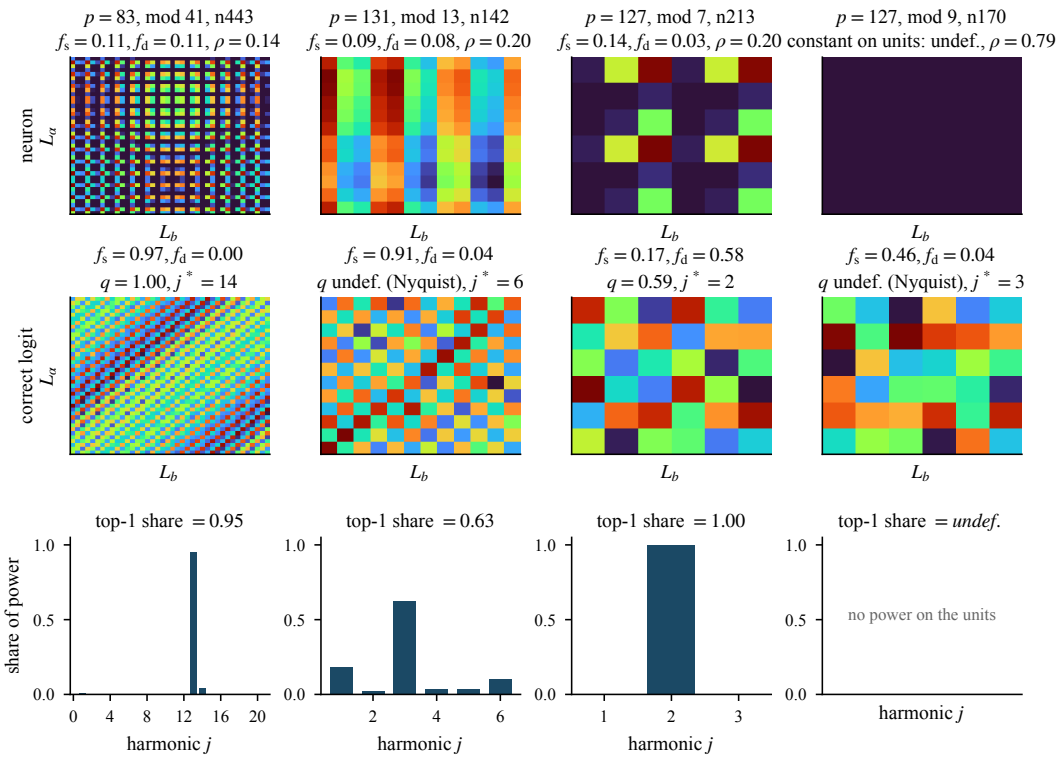


Figure 16: Clock diagnostics per cyclic component in the doubly-logged coordinate (L_α, L_b). Row 1: the top band neuron. Row 2: the logit of the correct answer, where the difference terms have cancelled. Row 3: the harmonic spectrum of the residue-pooled neuron response. The mod-9 column is blank by construction, see the text.

H THE READ-OUT TEST, AND ONE STATISTIC THAT IS TAUTOLOGICAL

Figure 17 gives the restriction matrices. Rows are what was kept in the MLP output at the answer position, columns the decoded component.

What is not reported, and why. The natural statistic here is "are the band's logits a function of $\gamma \bmod m$ alone". That statistic is *tautological*. The band basis is built from $\{\cos, \sin\}(2\pi k\gamma/q)$ for $k \in F_m$ evaluated over the dataset, so projecting the activations onto it forces every neuron's column into the span of those characters, and any function on $\mathbb{Z}/q\mathbb{Z}$ supported on F_m is a function of $\gamma \bmod m$. We confirmed the consequence: on a randomly initialized network the corresponding matrix is the identity to three decimals, diagonal 1.000 and off-diagonal 0.000. It measures the projection, not the network, and we report the decode instead, which must pass through W_U and an argmax over answer tokens.

What the decode shows. Restricting to the union of the CRT bands leaves joint accuracy at 0.966 to 1.000 across the nine models, against 0.003 to 0.018 for the randomly initialized control. Restricting to a single band F_m leaves A_m at 0.976 or above while every other component drops to its baseline chance. Restricting to the non-CRT control band leaves every component at its baseline.

One consistency check comes free. When only band F_m is kept, the joint accuracy should equal the product of the per-component decodes if the read-out treats the components as independent factors, since the surviving component is then correct always and the others only at their majority rate. The largest deviation from that product over the nine models is 0.036.

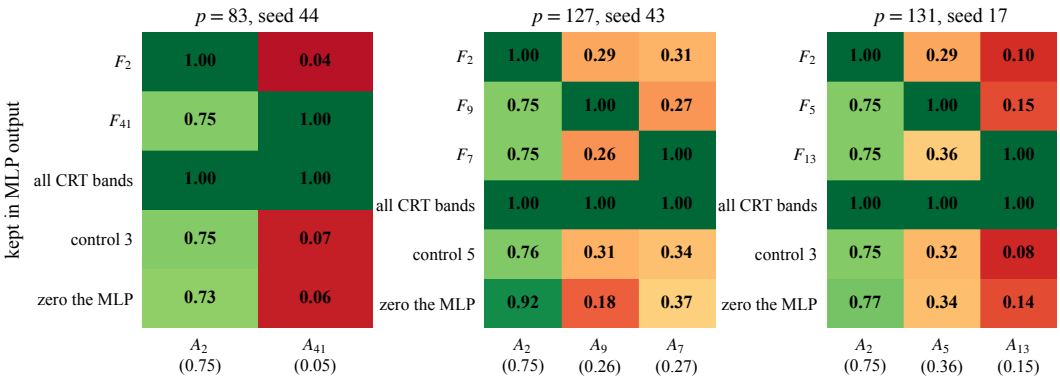


Figure 17: Restriction: keep only the named subspace of the MLP output and decode every component. The row that carries the claim is "all CRT bands".

I THE PERMUTATION NULL

The control for a prominence ratio is a random token order. A single random order, which is what the earlier version of this analysis used, cannot say whether an observed ratio is unusual, so we draw 200 orders per map and report the resulting null distribution. The p-value is the add-one estimator $p = (1 + \#\{\text{null} \geq \text{observed}\})/(n + 1)$, which is bounded below by $1/201 \approx 0.005$ rather than reaching zero; a reported $p \leq 0.005$ therefore means no null draw matched the observation. The discrete-log order is the observed statistic and is not part of the null sample. The discrete-log order gives 1.59 to 2.57 for W_E and 1.95 to 5.40 for W_U, which is 14 to 85 null standard deviations above the null mean and above every draw in all nine models, an empirical $p \leq 0.005$ at this sample size. For the CRT-aligned heads: 18 of 36 heads in total, each at $p \leq 0.005$, while every one of the other 18 sits at $p \geq 0.025$.

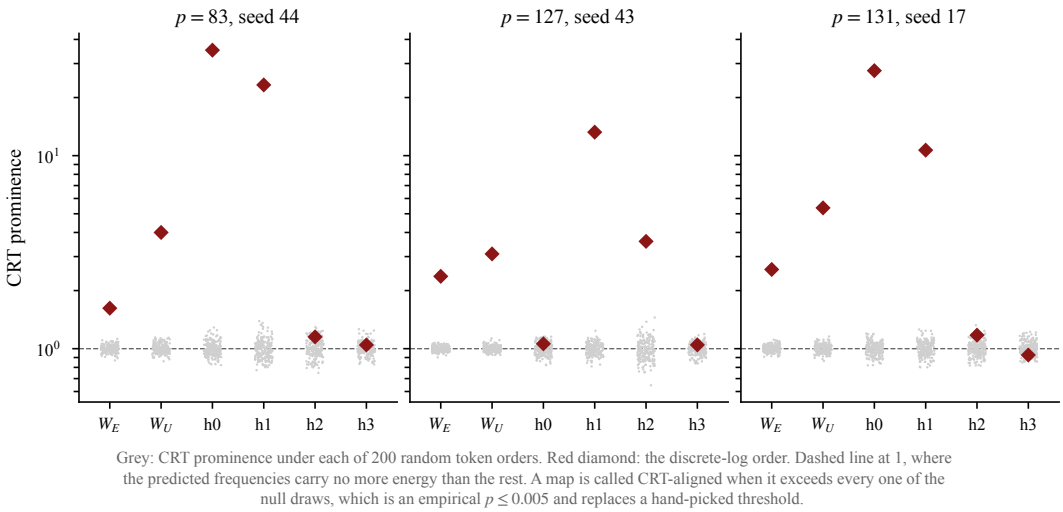


Figure 18: CRT prominence under 200 random token orders (grey) against the discrete-log order (red). Note the log scale: the null sits on 1 with standard deviation 0.03 to 0.11, and the aligned maps are one to two orders of magnitude above it.

J WHERE THE CAUSAL TESTS DEPART FROM THE PREDICTION

Five departures show up across the nine models. All are visible in Figures 19 and 20. We group them by kind, because two distinct things are happening: a band whose removal takes its component *below* chance, and a band that is not the sole carrier of its component.

Three components fall below their baseline chance. Ablating the mod-9 band of $p = 127$ leaves A_9 at 0.132, 0.080 and 0.156 on seeds 43, 45 and 46, against a baseline chance of 0.259; ablating the mod-7 band of the same prime leaves A_7 at 0.190 on seed 45, against 0.265; and ablating the mod-41 band of $p = 83$ leaves A_{41} at 0.022 on seed 44, against 0.048. A component that has merely lost its information should land on its baseline, not beneath it, so in these cases the intervention is not only deleting the residue but pushing the prediction toward a wrong one.

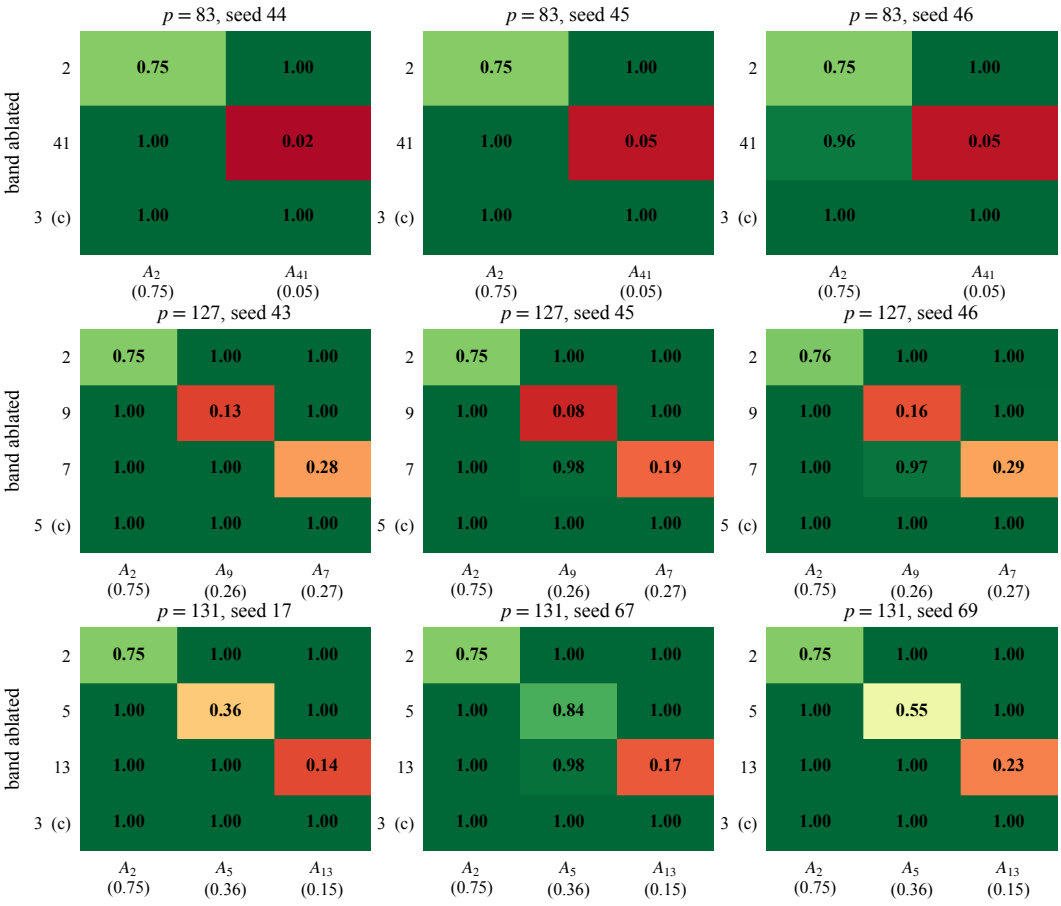
This also contradicts an assumption stated in Appendix B, that an ablated model's output distribution converges to the label prior. Below-baseline decodes mean it does not, at least for these components: removing a subspace is not the same as removing information, because the projection also perturbs what remains. The majority-residue rate is therefore the baseline chance a model reaches when it has *only* lost the residue, and is better read as a reference point than as a hard lower bound. We do not have an account of why these components in particular overshoot it.

Three components resist ablation, staying above their baseline. For mod-5 of $p = 131$, A_5 drops to 0.357 on seed 17, which is its baseline chance of 0.360, but only to 0.546 on seed 69 and 0.840 on seed 67. For mod-13 of the same prime, A_{13} drops to 0.144 on seed 17, at its baseline of 0.148, but only to 0.170 on seed 67 and 0.231 on seed 69. For mod-7 of $p = 127$, A_7 ends at 0.277 and 0.286 on seeds 43 and 46 against a baseline of 0.265, which is at baseline rather than above it. In the cases that stay clearly above, the band is not the sole carrier of that residue in that model. This is the sense in which the independence of the streams is approximate rather than exact, and the abstract's claim that each band controls its own stream should be read as holding cleanly on the large majority of component-seed pairs rather than on all of them.

The same model leaks under patching. On $p = 131$ seed 67, patching the mod-13 band drops the mod-5 decode to 0.678, the lowest off-diagonal entry anywhere in the nine models; the next lowest is 0.953. Combined with the weak ablation above and with the weakest sufficiency diagonal in the set (0.762 for mod-5), this is one phenomenon rather than three: on that seed the mod-5 and mod-13 streams are partly entangled rather than independent, so an intervention on either moves the other. The mechanism is clean on the other eight models and we report this one as a genuine exception.

K CAUSAL INTERVENTIONS ACROSS EVERY SEED

The main-body matrices show one model per prime. These are all nine, one row per prime and one column per seed, and they are where the seed-to-seed spread lives. Two cases the main text calls out are visible here: the mod-9 component of $p = 127$ lands below its baseline chance in all three seeds, and the mod-5 component of $p = 131$ falls to 0.357 on seed 17 but only to 0.546 and 0.840 on seeds 69 and 67.



Floor in brackets is the majority-residue rate; (c) marks the non-CRT control band.

Figure 19: Necessity, all nine models.

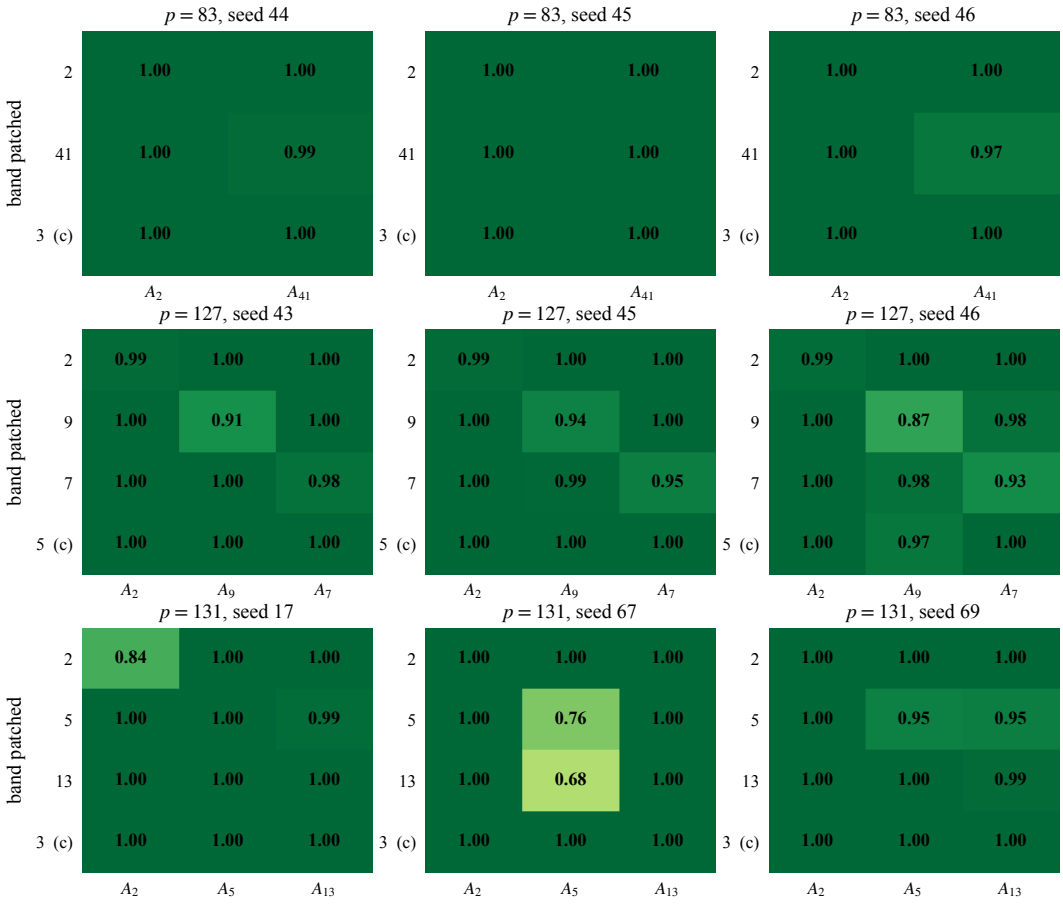


Figure 20: Sufficiency, all nine models.

L THE PRIME-POWER FILTRATION IN DETAIL

Why the characters inherit the filtration. The quotient maps of Equation 6 are reduction modulo successive powers of ℓ, so each term is a coarsening of the one to its left. Dually, a character of $\mathbb{Z}/\ell^t\mathbb{Z}$ pulls back along the reduction $\mathbb{Z}/\ell^e\mathbb{Z} \twoheadrightarrow \mathbb{Z}/\ell^t\mathbb{Z}$ to a character of $\mathbb{Z}/\ell^e\mathbb{Z}$ whose order still divides ℓ^t. The characters of order dividing ℓ^t therefore form a subgroup, and these subgroups nest,

$$1 \subset \widehat{\mathbb{Z}/\ell\mathbb{Z}} \subset \widehat{\mathbb{Z}/\ell^2\mathbb{Z}} \subset \cdots \subset \widehat{\mathbb{Z}/\ell^e\mathbb{Z}}, \tag{10}$$

an increasing chain dual to the decreasing chain of quotients. The relation between orders is containment, not interaction. Concretely, writing the residue in base ℓ as $\gamma \bmod \ell^e = d_0 + \ell d_1 + \cdots + \ell^{e-1} d_{e-1}$, a character of order ℓ^t sees only the bottom t digits.

What the ablation predicts, digit by digit. Ablating the characters of order exactly ℓ^t should break digit d_{t-1} and leave every lower digit intact. The decodes then inherit the damage by conjunction, since A_{ℓ^s} requires $d_0, \ldots, d_{s-1}$ to be simultaneously correct. Decoding the digits separately confirms this directly.

Table 8: Per-digit decode under each order ablation, $p = 113$ GELU, mod-16 component. The bold diagonal is the digit each order carries. Every digit below the ablated order survives at 0.9 or above; the A_{2^s} columns are the conjunctions of the digits to their left.

ablated order	rank	d_0	d_1	d_2	d_3	A_2	A_4	A_8	A_{16}
none	—	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
order-2 $\{56\}$	1	**0.920**	0.956	0.953	0.964	0.920	0.920	0.920	0.920
order-4 $\{28\}$	2	1.000	**0.876**	0.876	1.000	1.000	0.876	0.876	0.876
order-8 $\{14, 42\}$	4	0.999	0.983	**0.631**	0.650	0.999	0.983	0.622	0.528
order-16 $\{7, 21, 35, 49\}$	8	1.000	0.999	1.000	**0.429**	1.000	0.999	0.999	0.428
full band	15	0.696	0.541	0.568	0.554	0.696	0.393	0.256	0.184
control 3 (non-CRT)	—	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
baseline chance						0.750	0.500	0.312	0.188

Breadth and depth are independent. A low-order ablation touches more of the answer, because every residue depends on d_0, but it touches it less, because a projection can remove only as much as the subspace it removes spans. The order-2 level is a single frequency at the Nyquist bin $q/2$, where $\sin(2\pi k\gamma/q)$ vanishes identically, so it contributes one direction rather than two. Severity tracks rank exactly: 0.920, 0.876, 0.631, 0.429 for ranks 1, 2, 4, 8.

The same filtration with $\ell = 3$. The mod-9 component of $p = 127$ has two orders rather than four: the frequency 42 has order 3 and $\{14, 28, 56\}$ have order 9. Removing the order-3 character takes A_3 and A_9 down together, to 0.773 and 0.773 on seed 43, and to the same pair of equal values on the other two seeds. Removing the order-9 characters leaves A_3 at 1.000 and takes A_9 to 0.157. The asymmetry is the filtration. That $\gamma \bmod 3$ is a quantity the model represents, and not an artifact of our decoder, is confirmed by a probe: for $p = 127$ the mod-3 probe reaches 100.0% at the read-out against a random-label control of 51.6%, while for $p = 83$ and $p = 131$, where $3 \nmid q$, the same probe sits between 31% and 40% against a baseline of 33%.

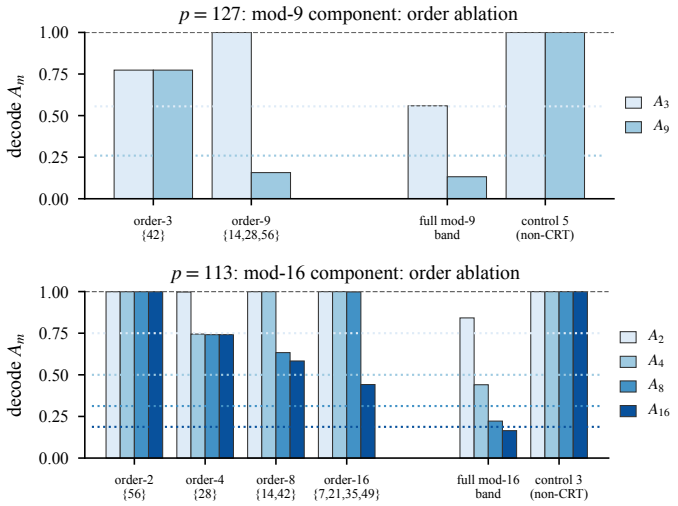


Figure 21: Top: the mod-9 component of $p = 127$, two orders. Bottom: the mod-16 component of $p = 113$ in the ReLU model, where the staircase matches the GELU model of Figure 7 except that the order-2 level, spanning a single direction, removes nothing measurable at all.

M TRAINING DYNAMICS AT THE OTHER PRIMES

The same normalized weight-norm measure on the other two trajectories. Only $p = 83$ uses ReLU; the other two are GELU models and are not among the nine analyzed elsewhere.

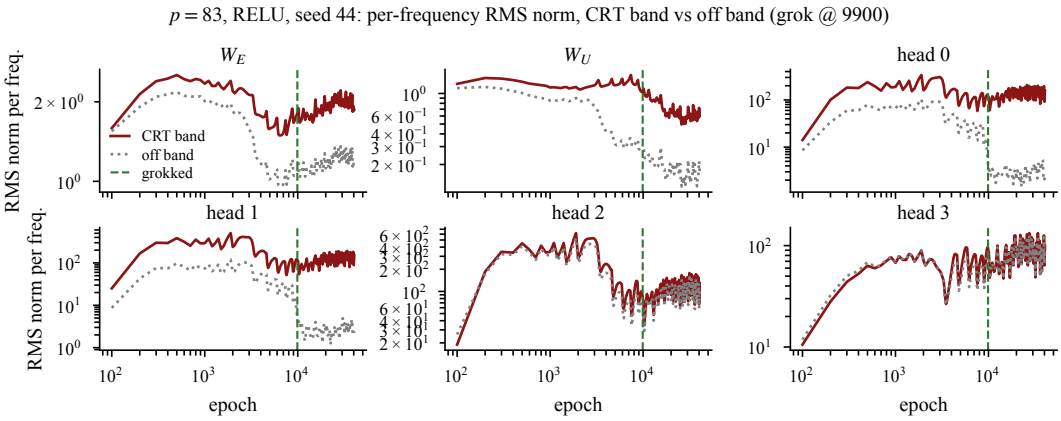


Figure 22: $p = 83$ ReLU seed 44. Heads 0 and 1 separate by roughly a factor of 60 between the two bands by the end of training, while heads 2 and 3 never separate. This is the longest trajectory in the set, running to epoch 40,500 against a grok epoch of 9900, so it shows the cleanup phase most completely.

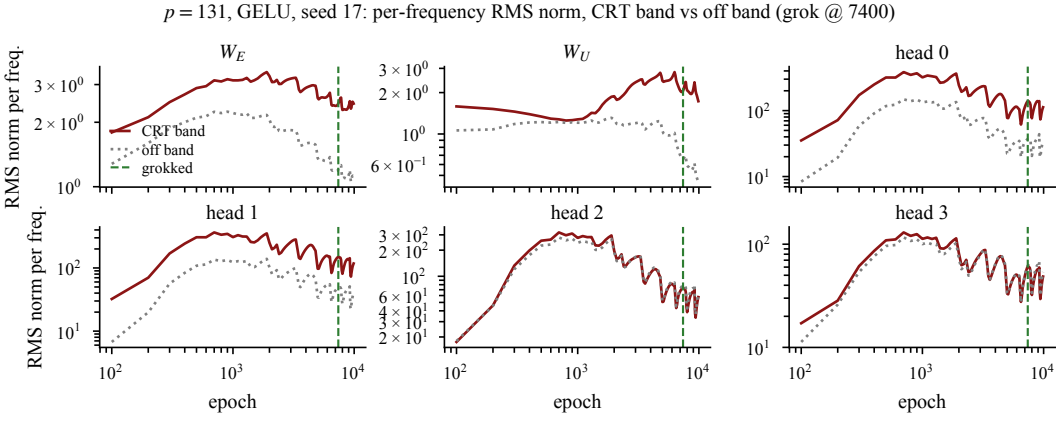


Figure 23: $p = 131$ GELU seed 17. The same two-good-two-flat pattern appears, with heads 0 and 1 separating and heads 2 and 3 staying together, but at smaller magnitude. This trajectory reaches 100% at epoch 7400 and the kept snapshots stop at 8900, so it covers only the first 1500 epochs of cleanup and should be read as under-observed rather than as a negative result.

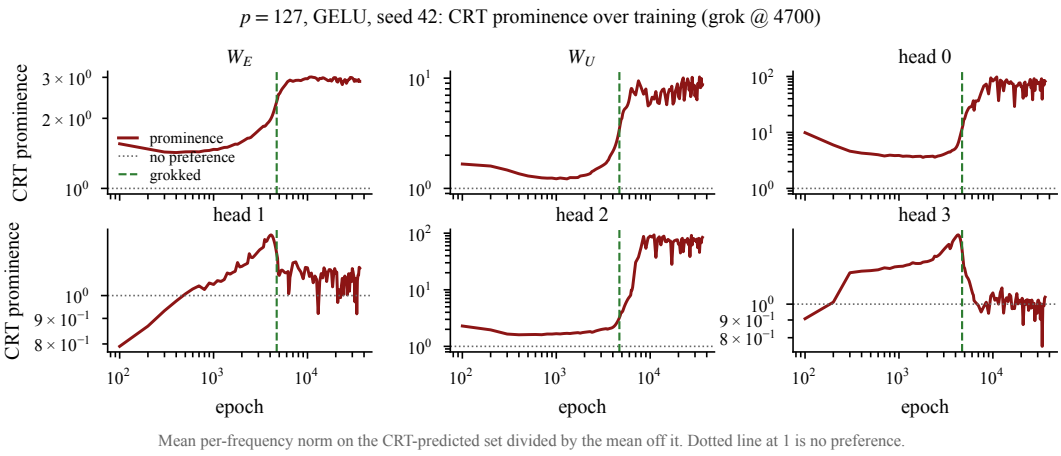


Figure 24: The same trajectory as Figure 8 summarized as a single ratio per structure: the mean per-frequency norm on the CRT band divided by the mean off it. This is the statistic the converged-weight analysis reports, tracked from the first snapshot. It is a ratio, so it cannot distinguish the CRT band growing from the off band shrinking; the two-curve figures above are what separate those.