From Groups to Rings: Causal Evidence for Algebraic Decomposition in Grokked Transformers

Shurui Zheng[1,2,3*] **Fanhong Li**[3]
Zhixing Huang[2] **Zixi Li**[2] **Lei Ji**[1]
[1]Microsoft Research Asia [2]Qiuzhen College, Tsinghua University
[3]Department of Psychology and Cognitive Science, Tsinghua University

Abstract

Prior mechanistic accounts of grokking have identified discrete Fourier transforms, group-character representations, and Chinese Remainder Theorem circuits. We connect these accounts through a formal Wedderburn-alignment framework based on classical algebra. Experiments cover 11 commutative finite algebras, including six with nontrivial Jacobson radical, and selected non-commutative tasks. Random-subspace controls and above-chance non-target preservation on base-error examples support component-organized representations beyond task accuracy alone. Cross-fitted post-attention interventions in three representative models provide stronger causal evidence for several components, while other components remain weak or probe-dependent. In a radical-containing example, coefficient interventions identify a low-dimensional output mediator but leave the internal multiplication algorithm open. Non-commutative group multiplication exhibits selective alignment; additional $\mathbb{F}_2[D_5]$ runs reveal a capacity–optimization interaction. These results support algebra-component-aligned representations and component-specific causal effects in the tested small Transformers, without establishing complete, probe-independent computational decomposition or a general theory of grokking dynamics.

1 Introduction

Grokking—the phenomenon of delayed generalization long after memorization—was discovered by Power et al. [2022] on modular arithmetic tasks. Subsequent mechanistic work has identified the algorithms that grokked networks learn: Nanda et al. [2023] demonstrated discrete Fourier transform circuits in Transformers trained on $\mathbb{Z}_{113}$ addition, and Chughtai et al. [2023] showed group character representations for group multiplication. Morwani et al. [2024] provided a theoretical account showing that max-margin solutions under weight decay necessarily align with Fourier features for cyclic groups. But a deeper question remains: *what structural transformation does a network undergo during grokking, and what determines whether it succeeds?*

These prior results share a limitation: they are stated in terms specific to the algebraic structure at hand—DFT for cyclic groups, GCR for finite groups—without a unifying language. The DFT/GCR accounts above are most directly connected to semisimple representation theory. Radical-containing algebras require additional structure, and non-commutative algebra multiplication must be distinguished from multiplication of group elements.

Contributions. We make three contributions:

*Corresponding author: zhengsr23@mails.tsinghua.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

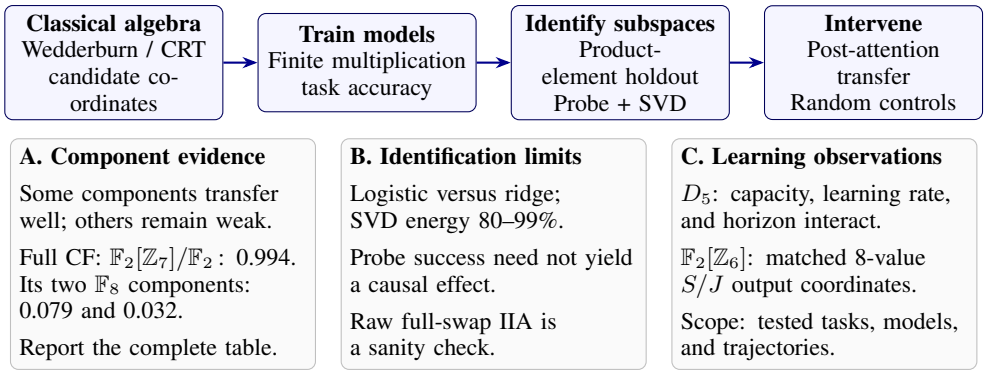


Figure 1: Framework and evidence. Classical algebra supplies candidate coordinates; neural subspaces and their causal effects are empirical questions. Full-state transport is a sanity check, whereas cross-fitted component interventions and controls test more specific hypotheses. The complete mixed results, probe sensitivity, corrected training dynamics, and D_5 configurations are reported below and in the appendices. No uniform layer-wise or probe-independent direct sum is inferred.

1. **Unified framework.** We define *Wedderburn-aligned representations* (Definition 1), subsuming DFT, GCR, and Chinese Remainder Theorem (CRT) circuits as special cases (Propositions 1–3). The propositions apply classical algebraic facts; our contribution is the formal interpretability connection and its empirical evaluation, rather than new algebraic theorems or guarantees about Transformer training.
2. **Causal evidence via IIA.** Using IIA [Geiger et al., 2024], we study component organization across 11 commutative algebras (5 semisimple, 6 radical-containing; char 2/3/5/7), $\mathbb{F}_2[S_3]$ algebra multiplication, and S_5 group multiplication. The evidence combines random-subspace controls, non-target preservation, and error compartmentalization with new cross-fitted post-attention interventions. The latter are strong for several components and weak or method-dependent for others; raw full-swap IIA is a descriptive sanity check.
3. **Grokking associations and dynamics.** A sweep across 12 algebras documents descriptive associations with algebraic structure. Additional $\mathbb{F}_2[D_5]$ runs show that its original failure is configuration-dependent. Corrected, equal-cardinality semisimple/radical output coordinates in $\mathbb{F}_2[\mathbb{Z}_6]$ reveal semisimple-first learning in three trajectories, replacing the earlier marginal-projection comparison.

2 Background

2.1 Relation to prior mechanistic accounts

DFT circuits for modular addition [Nanda et al., 2023] and group-character circuits for group multiplication [Chughtai et al., 2023] motivate the common componentwise description used here; CRT supplies its polynomial-ring counterpart. These connections use established algebra, not new versions of the decomposition theorems. Our formal definition organizes the comparison, while the empirical extension concerns radical-containing finite algebras and controlled component interventions. This differs from optimization analyses such as Morwani et al. [2024], which derive feature emergence under specified assumptions. We use the causal-abstraction methodology of Geiger et al. [2021, 2024] to test the proposed correspondences; the tests provide evidence at the measured locations rather than a theorem about training dynamics. Additional mechanistic and optimization context appears in Appendix O.

2.2 Wedderburn–Artin decomposition

Let k be a finite field and A a finite-dimensional k-algebra with Jacobson radical $\mathrm{rad}(A)$. The Artin–Wedderburn theorem gives $A/\mathrm{rad}(A) \cong \prod_{i=1}^{r} \mathrm{Mat}_{n_i}(D_i)$, where D_i are finite field extensions of k. When A is commutative, all $n_i = 1$. The Wedderburn–Malcev theorem ensures $A = S \oplus \mathrm{rad}(A)$

as k-vector spaces for some semisimple subalgebra S. We write $\pi_i : A \twoheadrightarrow \mathrm{Mat}_{n_i}(D_i)$ for the composed projections. Key examples: group algebras $k[G]$ (semisimple when $\mathrm{char}(k) \nmid |G|$ by Maschke's theorem) and polynomial quotient rings $k[x]/(f_1 \cdots f_r) \cong \prod_i \mathbb{F}_{q^{d_i}}$ (by the CRT, for distinct irreducible f_i). For repeated factors, CRT instead gives $k[x]/(\prod_i f_i^{e_i}) \cong \prod_i k[x]/(f_i^{e_i})$. Each local factor with $e_i > 1$ contains both a residue-field part and a radical. Independence between these direct-product factors is distinct from independence between S and $\mathrm{rad}(A)$ within a factor; the latter are coupled in multiplication.

2.3 Interchange intervention accuracy (IIA)

IIA [Geiger et al., 2021, 2024] tests whether an intermediate representation mediates a posited causal variable. Given a hypothesized component π_i, we select input pairs (a, a') with matching non-target components ($\pi_j(a) = \pi_j(a')$ for $j \neq i$) but differing target ($\pi_i(a) \neq \pi_i(a')$), replace the full position-0 activation from the source into the base forward pass, and check whether the output matches the counterfactual target $a' \cdot b$. For algebraic multiplication, this full embedding swap can reproduce an ordinary source-input evaluation, so its high accuracy alone does not identify internal component channels. We therefore distinguish this diagnostic from subspace-restricted interventions, use **error compartmentalization** (Section 4.4) as a complementary control, and test post-attention activations with held-out products (Appendix K).

3 Unified framework: Wedderburn-aligned representations

3.1 Definitions

Definition 1 (Wedderburn-aligned representation)**.** *Let A be a finite-dimensional k-algebra with a fixed Wedderburn–Malcev decomposition $A = S \oplus \mathrm{rad}(A)$, $S \cong \prod_{i=1}^{r} \mathrm{Mat}_{n_i}(D_i)$ (the choice of S is unique up to conjugacy by $1 + \mathrm{rad}(A)$). Let $\pi_i : A \twoheadrightarrow \mathrm{Mat}_{n_i}(D_i)$ denote the composed projections, $\pi_{\mathrm{rad}} : A \to \mathrm{rad}(A)$ the radical projection, and $\pi_{\mathrm{ss}} : A \twoheadrightarrow S$ the semisimple projection.*

A network $f_\theta : A \times A \to A$ with embedding $E : A \to \mathbb{R}^d$ learns a Wedderburn-aligned representation *if there exist subspace decompositions $\mathbb{R}^d = V_1 \oplus \cdots \oplus V_r \oplus V_{\mathrm{rad}}$ such that:*

- ***(WA1) Component encoding.*** *The embedding factorizes along Wedderburn components: $E(a) = (E_1(\pi_1(a)), \ldots, E_r(\pi_r(a)), E_{\mathrm{rad}}(\pi_{\mathrm{rad}}(a)))$.*
- ***(WA2) Causal independence (IIA form).*** *For each semisimple component i, the V_i-channel activation depends only on $\pi_i(a)$ and $\pi_i(b)$. A necessary intervention consistency check is that, for any $a, a', b, b' \in A$ with equal component-i operand coordinates, exchanging these channel activations preserves the corresponding computation. Passing matched-input or full-state tests alone is not sufficient to establish this dependency restriction.*

Definition 1 specifies the ideal organization being tested. Our measurements separately assess input encoding, causal effects at particular intermediate locations, and within-component algorithms; a successful test for one component does not certify a complete direct sum across all components and layers.

Remark 1 (Radical coupling)**.** *For non-semisimple A, the radical component of ab depends on semisimple components: $\pi_{\mathrm{rad}}(ab) = \pi_{\mathrm{rad}}(a)\pi_{\mathrm{ss}}(b) + \pi_{\mathrm{ss}}(a)\pi_{\mathrm{rad}}(b) + \pi_{\mathrm{rad}}(a)\pi_{\mathrm{rad}}(b)$. This coupling is compatible with independence between distinct CRT local factors. Error preservation between a local factor and a separate field factor (Section 4.5) tests that factorization; it does not demonstrate independence of the full S and $\mathrm{rad}(A)$ coordinates. We use separate equal-cardinality coordinates for their learning dynamics (Section 5.3).*

3.2 Prior circuits as special cases

We say a network additionally satisfies **(WC3)** if each component function computes the multiplication in its Wedderburn factor (formal definition in Appendix D). Propositions 1–3 record applications of classical Fourier, Artin–Wedderburn, and CRT facts to this description. They do not guarantee that gradient descent discovers the stated representations or that every trained model satisfies WA1–WA2 or WC3.

Proposition 1 (DFT circuits are Wedderburn circuits)**.** *Let* $\mathrm{char}(k) \nmid n$*. Then* $k[\mathbb{Z}_n] \cong k[x]/(x^n - 1) \cong \prod_{d|n} k[x]/\Phi_d(x)$*. When* k *contains all* n*-th roots of unity, each cyclotomic factor splits into 1-dimensional components indexed by characters* $\chi \in \widehat{\mathbb{Z}_n}$*, recovering the DFT:* $\widehat{f * g}(\chi) = \hat{f}(\chi)\hat{g}(\chi)$*. The resulting network is Wedderburn-aligned (WA1–WA2) with* $r = n$*, all* $n_i = 1$*,* $D_i = k$*,* $\mathrm{rad} = 0$*, and satisfies (WC3) since componentwise multiplication is scalar multiplication in* k*.*

Proposition 2 (GCR circuits are Wedderburn circuits)**.** *Let* $\mathrm{char}(k) \nmid |G|$*. The Wedderburn decomposition* $k[G] \cong \prod_i \mathrm{Mat}_{n_i}(D_i)$ *via irreducible representations* ρ_i *gives the GCR circuit of Chughtai et al. [2023]: embedding* $E_i(g) = \rho_i(g)$*, MLP computes* $\rho_i(a)\rho_i(b)$*, readout via character orthogonality* $\sum_i \frac{n_i}{|G|}\mathrm{tr}(\rho_i(ab)\rho_i(c^{-1})) = \mathbb{1}[ab = c]$*. This satisfies WA1–WA2 and WC3.*

Proposition 3 (CRT circuits are Wedderburn circuits)**.** *Let* $f = f_1 \cdots f_r \in k[x]$ *with distinct irreducible factors. Then* $A = k[x]/(f) \cong \prod_i \mathbb{F}_{q^{d_i}}$ *is semisimple with* $\mathrm{rad}(A) = 0$*. The CRT projections* $\pi_i : a \mapsto a \bmod f_i$ *are the Wedderburn projections, and multiplication in each* $\mathbb{F}_{q^{d_i}}$ *is the component computation (WC3).*

Proofs of Propositions 1–3 are in Appendix A.

4 Experiment 1: causal independence via IIA

We test component-organization hypotheses motivated by Definition 1 across multiple algebras, characteristics, and training phases, distinguishing embedding diagnostics from intermediate causal interventions.

4.1 Experimental setup

Model. One-layer Transformer (d_model=128, 4 heads, d_mlp=512, no LayerNorm) following Nanda et al. [2023]. Training: AdamW, lr=10^{-3}, weight decay=1.0, batch size=512, 100K steps. We use a training fraction of 0.3 (30% of the full multiplication table) for all algebras, except $|A| \leq 64$ where we use 0.5.

Intervention protocol. Full embedding swaps are descriptive sanity checks. Error-case preservation compares non-target prediction coordinates on base-error examples; the retained coordinate need not itself be wrong. Conditional IIA instead restricts preservation to base-correct records. For subspace tests, a logistic probe predicts component labels, SVD retains 95% of coefficient energy, and patching uses $h_{\mathrm{patch}} = h_b + VV^\top(h_s - h_b)$. The complete embedding protocol and metric definitions are in Appendix N.

Intermediate tests and code. We retrain $\mathbb{F}_2[\mathbb{Z}_7]$, $\mathbb{F}_3[\mathbb{Z}_4]$, and $\mathbb{F}_2[\mathbb{Z}_6]$ (seed 42, 30% training split), identify post-attention subspaces on disjoint product-element partitions, and evaluate them against 20 same-dimensional random subspaces. Appendix K gives full results and probe/SVD sensitivity. Code for these representative experiments is at `https://github.com/Ziegel-Thu/from-groups-to-rings`; it does not cover every historical table or checkpoint.

Algebras tested. Table 1 lists all algebras with their Wedderburn decompositions.

4.2 Four-level baseline hierarchy

To establish that IIA reflects genuine algebraic structure rather than artifacts of high accuracy or model architecture, we test four levels of control on $\mathbb{F}_2[x]/(\Phi_7)$:

The grokked model's IIA is 0.998, compared with 0.564 for the memorized model and 0.129 ± 0.010 for same-dimensional random subspaces. The comparison shows that the chosen intervention directions matter beyond the model's task accuracy. A local-ring control, $\mathbb{F}_2[x]/(x^7)$, has no nontrivial CRT product decomposition and yields random-subspace IIA 0.029 ± 0.006. Two additional embedding controls give CRT-subspace IIA 0.992 versus random 0.245 ± 0.008 for $\mathbb{F}_2[\mathbb{Z}_7]$, and 0.801 versus 0.129 for $\mathbb{F}_2[x]/(x^4(x^3 + x + 1))$. The latter gap supports selective causal effects under the identified subspaces; the loss relative to full-swap IIA does not by itself identify nonlinear encoding or its algorithm. Non-target preservation (NTP) supplies a complementary output diagnostic (Appendix C).

Table 1: Algebras tested with IIA and their Wedderburn–Artin decompositions. Additional algebras from the grokking sweep are listed in Appendix B. †Non-commutative. ‡Group multiplication task.

Algebra	$\lvert A\rvert$	char	$\dim \mathrm{rad}/\dim A$	Loewy	Semisimple quotient
$\mathbb{F}_2[x]/(\Phi_7) \cong \mathbb{F}_8 \times \mathbb{F}_8$	64	2	0	1	$\mathbb{F}_8 \times \mathbb{F}_8$
$\mathbb{F}_2[\mathbb{Z}_7] \cong \mathbb{F}_2 \times \mathbb{F}_8 \times \mathbb{F}_8$	128	2	0	1	$\mathbb{F}_2 \times \mathbb{F}_8 \times \mathbb{F}_8$
$\mathbb{F}_2[\mathbb{Z}_6] \cong \mathbb{F}_2[\varepsilon]/(\varepsilon^2) \times \mathbb{F}_4[\varepsilon]/(\varepsilon^2)$	64	2	1/2	2	$\mathbb{F}_2 \times \mathbb{F}_4$
$\mathbb{F}_2[x]/(x^4(x^3+x+1))$	128	2	3/7	4	$\mathbb{F}_2 \times \mathbb{F}_8$
$\mathbb{F}_2[x]/(x^4(x^2+x+1))$	64	2	1/2	4	$\mathbb{F}_2 \times \mathbb{F}_4$
$\mathbb{F}_2[x]/(x^3(x+1)(x^2+x+1))$	64	2	1/3	3	$\mathbb{F}_2 \times \mathbb{F}_2 \times \mathbb{F}_4$
$\mathbb{F}_3[x]/(x^3(x-1))$	81	3	1/2	3	$\mathbb{F}_3 \times \mathbb{F}_3$
$\mathbb{F}_5[x]/(x^2(x-1))$	125	5	1/3	2	$\mathbb{F}_5 \times \mathbb{F}_5$
$\mathbb{F}_3[\mathbb{Z}_4] \cong \mathbb{F}_3 \times \mathbb{F}_3 \times \mathbb{F}_9$	81	3	0	1	$\mathbb{F}_3 \times \mathbb{F}_3 \times \mathbb{F}_9$
$\mathbb{F}_5[\mathbb{Z}_3] \cong \mathbb{F}_5 \times \mathbb{F}_{25}$	125	5	0	1	$\mathbb{F}_5 \times \mathbb{F}_{25}$
$\mathbb{F}_7[\mathbb{Z}_3] \cong \mathbb{F}_7 \times \mathbb{F}_7 \times \mathbb{F}_7$	343	7	0	1	$\mathbb{F}_7^3$
$\mathbb{F}_7[S_5]^{\dagger\ddagger}$	120	7	0	1	$\mathbb{F}_7^2 \times \mathrm{Mat}_4(\mathbb{F}_7)^2 \times \mathrm{Mat}_5(\mathbb{F}_7)^2 \times \mathrm{Mat}_6(\mathbb{F}_7)$

Table 2: Four-level baseline hierarchy on $\mathbb{F}_2[x]/(\Phi_7) \cong \mathbb{F}_8 \times \mathbb{F}_8$.

Condition	Train acc	Test acc	IIA	Description
Random init	0.014	0.016	0.020	Architectural bias only
Memorized (step 500)	0.997	0.098	0.564	Overfit, no generalization
Grokked (step 60K)	1.000	1.000	0.998	Full generalization
Random subspace (same dim)	1.000	1.000	0.129 ± 0.010	Patching random 2–3D subspace

4.3 Intermediate causal effects and probe dependence

At the post-attention answer token, cross-fitted logistic/95%-SVD subspaces yield exact full-counterfactual accuracy 0.994 for the $\mathbb{F}_2$ component of $\mathbb{F}_2[\mathbb{Z}_7]$, 0.737 for the first $\mathbb{F}_3$ component of $\mathbb{F}_3[\mathbb{Z}_4]$, and 0.564 for the 16-element local factor of $\mathbb{F}_2[\mathbb{Z}_6]$. However, the other $\mathbb{F}_2[\mathbb{Z}_7]$ components give only 0.079 and 0.032. The full table, including all weak components and random controls, appears in Appendix K.

Decodability does not guarantee a causal effect: the second $\mathbb{F}_3$ component has held-out probe accuracy 0.649, but target intervention accuracy 0.010 versus random 0.012. With SVD thresholds 80/90/95/99%, the strongest $\mathbb{F}_2[\mathbb{Z}_7]$ component stays at full-counterfactual accuracy 0.994; ridge gives 0.888. Other components vary substantially, and most lose causal efficacy with ridge. Cross-fitting mitigates probe-data reuse, but does not establish probe-independent subspaces. Weak results leave causal identification incomplete under the tested method. Full-state transport across embedding, post-attention, and post-MLP locations in the one-layer graph is mechanically expected and remains a sanity check, not additional evidence of component separation.

4.4 Multi-algebra IIA: semisimple case

Table 3 presents IIA results across semisimple algebras in characteristics 2, 3, 5, and 7.

Table 3: Original reported embedding diagnostics for semisimple algebras. Raw full-swap IIA is a sanity check. "Err. pres.": non-target prediction-coordinate preservation on examples where the base full-product prediction is wrong (chance in parentheses).

Algebra	char	Raw IIA	Err. pres.	Grok	Components
$\mathbb{F}_2[x]/(\Phi_7)$	2	.998	.724 (.125)	9/10	$\mathbb{F}_8 \times \mathbb{F}_8$
$\mathbb{F}_2[\mathbb{Z}_7]$	2	.996	.683 (.250)	10/10	$\mathbb{F}_2 \times \mathbb{F}_8 \times \mathbb{F}_8$
$\mathbb{F}_3[\mathbb{Z}_4]$	3	.996	.730 (.259)	10/10	$\mathbb{F}_3 \times \mathbb{F}_3 \times \mathbb{F}_9$
$\mathbb{F}_5[\mathbb{Z}_3]$	5	.989	.539 (.120)	7/10	$\mathbb{F}_5 \times \mathbb{F}_{25}$
$\mathbb{F}_7[\mathbb{Z}_3]$	7	.995	.663 (.143)	10/10	$\mathbb{F}_7 \times \mathbb{F}_7 \times \mathbb{F}_7$

Key finding: Above-chance error preservation across these semisimple tasks supports component-organized behavior across four tested characteristics. High raw IIA is consistent with their task accuracy; intermediate separation is assessed separately in Section 4.3.

Modular negative control. $\mathbb{F}_3[\mathbb{Z}_3]$ (27 elements, char 3 divides $|\mathbb{Z}_3| = 3$) is a local ring with radical fraction 2/3. It fails to grok in the original configurations (0/10 seeds, test accuracy ≤ 0.40), whereas $\mathbb{F}_3[\mathbb{Z}_4]$ groks at 10/10 seeds. Characteristic alone does not explain this contrast; size and other task properties also differ, so it does not isolate semisimplicity as a causal prerequisite.

Ruling out accuracy as a trivial explanation. Since raw IIA $\approx$ test accuracy for full-swap patching, high IIA alone does not prove structural decomposition. We additionally examine *error compartmentalization*: on examples where the base full-product prediction is wrong, do non-target projections of that prediction remain unchanged after intervention? The "Err. pres." columns in Tables 3 and 4 give preservation above chance across 11 tasks. This is evidence about error structure beyond high accuracy. The historical $\mathbb{F}_2[\mathbb{Z}_6]$ row uses marginal projections, so the collection is not a verification of complete CRT separation for every task.

4.5 Extending IIA to algebras with radical

We test IIA on six commutative algebras with nontrivial radical, spanning radical fractions 1/3–1/2 and characteristics 2, 3, and 5 (Table 4).

Table 4: Original embedding diagnostics for radical-containing algebras, averaged across grokked seeds. Raw columns denote the first and remaining measured projections (q_1, q_{rest}); for polynomial rings these are local/field CRT factors, while the $\mathbb{F}_2[\mathbb{Z}_6]$ row uses historical $\mathbb{Z}_2/\mathbb{Z}_3$ marginals. They are not full J/S coordinates. "Err." is error preservation (chance in parentheses).

Algebra	$\lvert A\rvert$	char	rf	Raw q_1	Raw q_{rest}	Err. (chance)
$\mathbb{F}_2[\mathbb{Z}_6]$	64	2	1/2	0.995	0.996	.643 (.188)
$\mathbb{F}_2[x]/(x^4(x^3+x+1))$	128	2	3/7	0.992	0.991	.454 (.094)
$\mathbb{F}_2[x]/(x^4(x^2+x+1))$	64	2	1/2	0.998	0.998	.303 (.156)
$\mathbb{F}_2[x]/(x^3(x+1)(x^2+x+1))$	64	2	1/3	0.996	0.997	.745 (.292)
$\mathbb{F}_3[x]/(x^3(x-1))$	81	3	1/2	0.995	0.995	.469 (.185)
$\mathbb{F}_5[x]/(x^2(x-1))$	125	5	1/3	0.969	0.973	.440 (.120)

Directional error analysis (polynomial quotient rings). We split error preservation by CRT factor type—local (radical-containing) versus field—for the polynomial quotient rings in the directional analysis (Table 5). A local factor includes its own residue-field coordinate; it is not the radical alone.

Table 5: Original reported directional preservation for polynomial quotient rings. "Patch $\rightarrow$ Preserve": we patch the first factor type and test whether the second type's prediction coordinate is preserved on base-error examples.

Direction	Err. pres.	Chance	Ratio
Local $\rightarrow$ field	0.494	0.195	2.5$\times$
Field $\rightarrow$ local	0.287	0.061	4.7$\times$

Both directions preserve errors above chance, supporting separation between these CRT factors. This diagnostic does not identify the within-factor radical multiplication algorithm. For $\mathbb{F}_2[\mathbb{Z}_6]$, the corrected CRT factors have 4 and 16 elements and both contain radical; their intermediate results are reported in Appendix K.

4.6 Non-commutative groups

For $\mathbb{F}_2[S_3]$ (non-commutative, modular, 64 elements), grokked models at d_model=256 exhibit selective embedding-subspace effects for the probed trivial and standard quotient blocks, with cross-block preservation ≥ 0.999 (Table 6). Dimensions 3 and 12 refer to neural subspaces, not algebraic block dimensions. These tests do not establish complete downstream alignment including the radical.

For S_5 group multiplication over $\mathbb{F}_7$ (semisimple, 7 irreps including matrix algebras up to $\text{Mat}_6(\mathbb{F}_7)$), 16/20 seeds grok, but models exhibit *selective* Wedderburn alignment rather than using all components.

Table 6: $\mathbb{F}_2[S_3]$ subspace IIA (d_model=256, 1-layer). Patching one block's subspace selectively affects that block only.

Swap subspace	dim	Target-block IIA	Other-block preserve
Trivial block ($\mathbb{F}_2$)	3	0.959	0.999
Standard block ($\mathrm{Mat}_2(\mathbb{F}_2)$)	12	0.999	1.000

Setup. $\mathbb{F}_7[S_5]$ is semisimple with 7 irreducible representations: $\mathbb{F}_7[S_5] \cong \mathbb{F}_7^2 \oplus \mathrm{Mat}_4(\mathbb{F}_7)^2 \oplus \mathrm{Mat}_5(\mathbb{F}_7)^2 \oplus \mathrm{Mat}_6(\mathbb{F}_7)$. Since S_5's irreducible representations have the same dimensions over $\mathbb{F}_7$ and $\mathbb{R}$, the network—which operates with real-valued weights—learns real representations whose dimensions match these components.

Selective alignment. Each grokked seed concentrates on 1–2 irreducible representations (Figure 2), with the dominant irrep approximately uniformly distributed across $\dim \geq 4$ components: $(2,1^3)$ (Mat_4) 5 times, $(3,1^2)$ (Mat_6) 5 times, $(4,1)$ (Mat_4) 3 times, $(3,2)$ (Mat_5) 3 times. One-dimensional irreps (trivial and sign) have FVE $= 0$ across all 20 seeds.

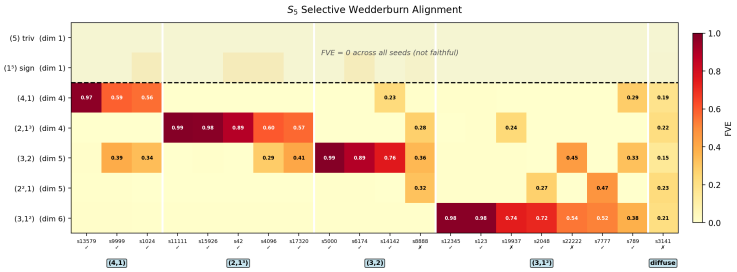


Figure 2: FVE across 20 seeds of S_5 group multiplication, grouped by dominant irrep. *Above dashed line:* the two dim=1 representations (trivial and sign) have FVE $= 0$ across all seeds. *Below:* among dim≥4 irreps, each grokked seed (✓) concentrates on a different 1–2 components (block-diagonal pattern); non-grokked seeds (×) remain diffuse.

Interpretation: faithfulness and weight decay. For group multiplication (as opposed to algebra multiplication), a single faithful irreducible representation suffices to distinguish all group elements. Since S_5's only proper nontrivial normal subgroup is A_5, any $\dim \geq 4$ irrep is faithful. A weight-decay preference for economical solutions is consistent with concentration on a sufficient faithful representation. In Table 7, eff_num decreases from 1.85 to 1.10 as d_model increases from 128 to 512, and each grokked d512 seed has one component above the 0.3 threshold. This is a concentration measurement, not exact absence of all other components, and the capacity sweep does not isolate weight decay as its cause. The trivial and sign irreps, with kernels S_5 and A_5, cannot individually distinguish all group elements. Their observed lack of concentration is consistent with this task requirement.

Table 7: S_5 capacity scaling: selective alignment intensifies with capacity.

d_model	Params	Grok rate	eff_num	$n_{\text{above }0.3}$
128	∼230K	16/20	1.85 ± 0.71	1.4 ± 0.5
256	∼900K	7/10	1.47 ± 0.44	1.2 ± 0.4
512	∼3.6M	9/10	1.10 ± 0.05	1.0 ± 0.0

The reported high-FVE components have larger above-chance λ-IIA gaps, but some moderate-FVE components also exceed chance. The per-seed results are in Appendix I; no sharp FVE threshold is inferred. Since the task is multiplication of group elements, this selective alignment is not evidence that all algebra blocks are learned when all are necessary for full group-algebra multiplication.

5 Experiment 2: grokking associations and dynamics

5.1 Descriptive algebraic associations

We conducted a grokking sweep across 12 algebras with varying size, radical fraction, and commutativity (Table 8).

Table 8: Original grokking sweep. Default: d_model=128, 1-layer. "d256" indicates capacity-enhanced runs. The D_5 0/12 entry concerns the original 10^{-3} learning rate and 100K-step horizon; additional configurations are reported in Appendix L.

Algebra	$\|A\|$	rf	Comm.?	Grok rate	Avg step	Config
$\mathbb{F}_2[\mathbb{Z}_7]$	128	0%	✓	10/10	12K	default
$\mathbb{F}_2[\mathbb{Z}_3^2]$	512	0%	✓	10/10	17K	d256
$\mathbb{F}_2[\mathbb{Z}_6]$	64	50%	✓	4/10	20K	default
$\mathbb{F}_2[\mathbb{Z}_8]$	256	88%	✓	6/10	34.5K	d256
$\mathbb{F}_2[\mathbb{Z}_2^3]$	256	88%	✓	5/10	42K	default
$\mathbb{F}_2[\mathbb{Z}_4 \times \mathbb{Z}_2]$	256	88%	✓	3/10	50K	default
$\mathbb{F}_3[\mathbb{Z}_3]$	27	67%	✓	0/10	—	default
$\mathbb{F}_3[\mathbb{Z}_6]$	729	67%	✓	0/5	—	default
$\mathbb{F}_2[D_5]$	1024	10%	×	0/12	—	4 configs
$\mathbb{F}_2[S_3]$	64	17%	×	4/10	14K	default
$\mathbb{F}_2[S_3]$	64	17%	×	**6/10**	16K	d256
$\mathbb{F}_2[D_4]$	256	88%	×	0/10	—	default
$\mathbb{F}_2[Q_8]$	256	88%	×	0/10	—	default

Radical fraction summarizes the proportion of algebra dimensions in the Jacobson radical. The original observed grokked-run collection reported a modest Spearman association with grokking time ($\rho = 0.35$, $p = 0.001$, $n = 82$). This is descriptive: size, characteristic, commutativity, model capacity, training fraction, and component counts co-vary, and conditioning on runs that grok introduces selection bias. It does not isolate radical fraction as a causal factor or establish it as a stronger predictor than size. The additional D_5 runs are a separate study and are not folded into this original statistic. At fixed finite dimension, possible radical fractions are discrete. Polynomial quotient rings offer controlled comparisons at fixed cardinality, but such a sweep is not supplied here and remains necessary to isolate the association.

5.2 Capacity and optimization in non-commutative tasks

The original $\mathbb{F}_2[D_5]$ result was 0/12 across four configurations, whereas $\mathbb{F}_2[S_3]$ grokked at 4/10 seeds with $d_{\text{model}} = 128$ and at 7/10 with d256/L2 (Table 9). These observations motivate a configuration sweep, not a universal threshold law attributable to non-commutativity.

Table 9: $\mathbb{F}_2[S_3]$ capacity sweep ($|A| = 64$, rf = 17%, non-commutative, 10 seeds each). Grokking defined as best test accuracy ≥ 0.95 at any point during training.

Config	Params	Grok rate	Median grok step
d128, 1-layer	∼115K	4/10	14.1K
d256, 1-layer	∼460K	6/10	16.1K
d256, 2-layer	∼590K	**7/10**	5.4K

Additional D_5 experiments at learning rate 10^{-4} demonstrate that the original failure is not absolute. The d1024/L2 configuration groks for seeds 42, 123, and 789 at 25K, 34K, and 31K steps, with final test accuracy 0.9994, 0.9983, and 0.9998. The d512/L4 run reaches the threshold at 112K, beyond the original 100K horizon. In the seed-42 width/depth grid, d128 and d256 fail by 300K; d512 succeeds only at L4; and d1024 succeeds at L1/L2/L4. Complete configurations and outcomes are in Appendix L.

The evidence supports a capacity–optimization interaction in this task: reducing the learning rate does not make every size succeed, and increasing capacity at the original settings was not sufficient

in the submitted runs. The sweep changes several settings relative to the original experiments and does not identify non-commutativity as the cause of a universal capacity threshold. Our framework describes candidate structure of learned solutions; it does not predict which optimizer settings reach them. The new D_5 runs establish grokking, with their component alignment still unmeasured.

5.3 Training dynamics with matched semisimple/radical coordinates

The original $\mathbb{F}_2[\mathbb{Z}_6]$ dynamics compared a 4-valued $\mathbb{Z}_2$ marginal with an 8-valued $\mathbb{Z}_3$ marginal. These are not jointly injective CRT coordinates, and the former is not the full radical. Its step-1K statistic was raw target accuracy over 2,000 trials, not conditional IIA. It cannot establish radical-before-semisimple emergence.

We instead use a bijective $A = S \oplus J$ coordinate system, with $|S| = |J| = 8$, and verify reconstruction and multiplication over all 64 elements and 4,096 products. Table 10 reports the earliest checkpoint from which test balanced accuracy remains at least 0.8 for three consecutive saved checkpoints (details in Appendix M). These are output-coordinate prediction metrics, not intervention accuracies or a demonstration of internal circuit formation.

Table 10: Corrected $\mathbb{F}_2[\mathbb{Z}_6]$ output-coordinate dynamics (three retrained trajectories). Both S and J have eight values. Emergence uses sustained test balanced accuracy ≥ 0.8; grokking uses full-product test accuracy ≥ 0.95. "—" means not reached within 100K steps.

Seed	Grok step	S emergence	J emergence
42	24K	9K	12K
123	—	11K	25K
789	—	7K	8K

All three trajectories learn the semisimple output coordinate before the radical coordinate by this criterion. Raw test accuracy yields the same ordering (S: 8K/10K/6K; J: 11K/24K/8K). Only seed 42 crosses the grokking threshold. The observed order is consistent with radical multiplication depending on both semisimple and radical inputs, but three trajectories do not establish a general dynamical law.

6 Discussion

Max-margin generalization. Can the max-margin analysis of Morwani et al. [2024], which proves Fourier feature emergence for cyclic groups, be extended to general commutative semisimple algebras?

Non-commutative alignment. Selective alignment in S_5 group multiplication is consistent with faithful irreps sufficing for the task and with a weight-decay preference for economical solutions. This does not establish that weight decay uniquely causes the observed selection. Complete downstream alignment for non-commutative algebra multiplication remains open; the $\mathbb{F}_2[S_3]$ quotient-block tests and the new D_5 grokking results address only parts of this question.

Radical component algorithms. Component organization and within-component algorithm identification are distinct claims. Semisimple models can show high Fourier-logit similarity (Appendix D), while a corrected coefficient intervention for $\mathbb{F}_2[\varepsilon]/(\varepsilon^2)$ identifies a 4D post-attention mediator for the nilpotent output coordinate. Its unconditional target/full-counterfactual accuracies are 0.977/0.904, and the base-and-source-correct conditional values are 0.988/0.918. These effects strengthen the radical evidence but do not identify the separate multiplication terms or a unique truncated-convolution algorithm; radical WC3 remains open (Appendix K.1).

6.1 Limitations

1. **Contribution and scope.** The contribution is a formal interpretability framework and empirical evidence based on classical algebra. It is not a new algebra theorem or a general theory of grokking dynamics. Tasks are synthetic finite-algebra operations in small Transformers.
2. **Intermediate identification.** Embedding patching tests input encodings, and full-state transport need not reveal separated computation. The new post-attention evidence covers three one-layer

seed-42 models and is mixed across components. Full decomposition through every layer is not established.

3. **Probe dependence.** Cross-fitting and random controls reduce confounds, but several subspaces change markedly with probe family or SVD threshold. Weak interventions leave identification incomplete; even successful ones do not establish a unique, probe-independent direct sum.
4. **Non-commutative coverage and scale.** S_5 is a group task with selective alignment; $\mathbb{F}_2[S_3]$ tests selected quotient blocks. The 1,024-element D_5 algebra now groks at selected settings, but its alignment is untested. Most mechanistic experiments use 1–2 layers; the new capacity sweep extends to four layers without establishing deeper mechanistic generality.
5. **Radical structure.** CRT local-factor separation and the S/J split are different. Low-dimensional output-coordinate mediation does not identify a complete radical algorithm. High-radical local rings have only one CRT factor; their internal filtration requires different tests.
6. **Associations and trajectories.** The radical-fraction correlation is confounded and conditional on grokking. The corrected S/J dynamics cover three trajectories, only one of which groks. Neither establishes a general causal account of learning speed.
7. **Reproducibility coverage.** The public package runs representative training and post-attention interventions from scratch. It does not yet reproduce every original sweep, historical checkpoint, or table.

7 Conclusion

We introduced Wedderburn-aligned representations as a unified framework for how grokked Transformers learn algebraic multiplication, subsuming DFT and GCR circuits as special cases. Across the tested finite-algebra tasks, embedding controls and error preservation support component-organized representations. Cross-fitted post-attention interventions add causal evidence for several components, alongside substantial method dependence and weak components. Corrected radical-coordinate dynamics and successful D_5 runs refine the empirical picture. The results support the framework in the tested settings while leaving complete downstream identification, radical algorithms, and the optimization dynamics of alignment open.

Acknowledgments and Disclosure of Funding

This work received no external funding. The authors declare no competing interests.

References

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.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 2024.

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Depen Morwani, Benjamin L Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham M Kakade. Feature emergence via margin maximization: Case studies in algebraic tasks. *International Conference on Learning Representations (Spotlight)*, 2024.

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.

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

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. *arXiv preprint arXiv:2312.06581*, 2024.

Pascal Junior Tikeng Notsawo, Hattie Zhou, Mohammad Pezeshki, Irina Rish, and Guillaume Dumas. Predicting grokking long before it happens: A look into the loss landscape of models which grok. *arXiv preprint arXiv:2306.13253*, 2024.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 2023.

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

A Proof sketches for Propositions 1–3

Proof sketch of Proposition 1. Since $\mathrm{char}(k) \nmid n$, the polynomial $x^n - 1$ is separable (coprime with its derivative), so $k[\mathbb{Z}_n] \cong k[x]/(x^n - 1)$ is semisimple by Maschke's theorem. The factorization $x^n - 1 = \prod_{d|n} \Phi_d(x)$ into cyclotomic polynomials, combined with the Chinese Remainder Theorem (the Φ_d are pairwise coprime), gives the first isomorphism. When $\mu_n \subseteq k^\times$, each $\Phi_d(x)$ splits into linear factors $(x - \zeta)$ over k, one per primitive d-th root of unity. Each linear factor yields a 1-dimensional component indexed by the corresponding character $\chi : g \mapsto \zeta^g$. The CRT projection onto this component is exactly the DFT coefficient $\hat{f}(\chi) = \sum_g f(g)\chi(g)$, and componentwise multiplication $\widehat{f * g}(\chi) = \hat{f}(\chi)\hat{g}(\chi)$ follows from the ring homomorphism property of each projection. The readout is the inverse DFT. □

Proof sketch of Proposition 2. Since neural networks operate over $\mathbb{R}$, we state this proof for $k = \mathbb{R}$ (or any field of characteristic 0), where the trace sum takes real values and maximization is well-defined. (The experimental tasks are defined over $\mathbb{F}_q$, but the learned representations are real-valued.) By Maschke's theorem ($\mathrm{char}(k) \nmid |G|$), the group algebra $k[G]$ is semisimple. The Artin–Wedderburn theorem gives $k[G] \cong \prod_i \mathrm{Mat}_{n_i}(D_i)$, where the product runs over isomorphism classes of irreducible k-representations $\rho_i : G \to \mathrm{Mat}_{n_i}(D_i)$. The embedding $E_i(g) = \rho_i(g)$ and the multiplication $\rho_i(a)\rho_i(b) = \rho_i(ab)$ follow from the representation homomorphism. For the readout, the character orthogonality relations give $\sum_i \frac{n_i}{|G|}\mathrm{tr}(\rho_i(g)) = \delta_{g,e}$, from which $\sum_i \frac{n_i}{|G|}\mathrm{tr}(\rho_i(ab)\rho_i(c^{-1})) = \delta_{ab,c}$. (Here tr denotes the matrix trace; when $D_i \neq k$, the orthogonality relations require the reduced trace $\mathrm{Tr}_{D_i/k} \circ \mathrm{tr}$. In our experiments, all $D_i = k$.) This is exactly the indicator function for $ab = c$, so the logit $\ell(a, b, c) = \sum_i n_i \mathrm{tr}(\rho_i(ab)\rho_i(c^{-1}))$ is maximized at $c = ab$, implementing (WA1)–(WA2) and (WC3) with $\mathrm{rad} = 0$. □

Proof sketch of Proposition 3. Let $f = f_1 \cdots f_r$ with f_i distinct irreducible polynomials. Since the f_i are pairwise coprime, the CRT gives $k[x]/(f) \xrightarrow{\sim} \prod_i k[x]/(f_i)$ via $g \mapsto (g \bmod f_i)_i$. Each $k[x]/(f_i) \cong \mathbb{F}_{q^{d_i}}$ is a field (since f_i is irreducible with $d_i = \deg f_i$), so A is a product of fields, hence semisimple with $\mathrm{rad}(A) = 0$. The CRT projections $\pi_i : g \mapsto g \bmod f_i$ are the Wedderburn projections, and multiplication within each $\mathbb{F}_{q^{d_i}}$ provides the component computation. The CRT inverse—reconstructing g from $(\pi_1(g), \ldots, \pi_r(g))$ via the idempotents $e_i \in A$ satisfying $e_i \equiv 1 \pmod{f_i}$ and $e_i \equiv 0 \pmod{f_j}$ for $j \neq i$—serves as the readout map. □

B Wedderburn decompositions and grokking sweep

Table 11 lists all algebras used in the grokking sweep (Section 5) that are not in Table 1.

Table 11: Additional algebras from the grokking sweep (no IIA analysis). †Non-commutative.

Algebra	$\lvert A\rvert$	char	rf	Loewy	A/rad
$\mathbb{F}_2[x]/(x^7)$ (local)	128	2	6/7	7	$\mathbb{F}_2$
$\mathbb{F}_2[\mathbb{Z}_4 \times \mathbb{Z}_2]$	256	2	7/8	5	$\mathbb{F}_2$
$\mathbb{F}_2[\mathbb{Z}_2^3]$	256	2	7/8	4	$\mathbb{F}_2$
$\mathbb{F}_2[S_3]^\dagger$	64	2	1/6	2	$\mathbb{F}_2 \times \mathrm{Mat}_2(\mathbb{F}_2)$
$\mathbb{F}_2[D_4]^\dagger$	256	2	7/8	5	$\mathbb{F}_2$
$\mathbb{F}_2[Q_8]^\dagger$	256	2	7/8	5	$\mathbb{F}_2$
$\mathbb{F}_2[D_5]^\dagger$	1024	2	1/10	2	$\mathbb{F}_2 \times \mathrm{Mat}_2(\mathbb{F}_4)$
$\mathbb{F}_3[\mathbb{Z}_3]$ (modular)	27	3	2/3	3	$\mathbb{F}_3$

The D_4 and Q_8 rows share the listed radical fraction and Loewy length; these summary invariants do not imply isomorphism of their group algebras or explain their optimization behavior. For $\mathbb{F}_2[D_5] \cong \mathbb{F}_2[t]/(t^2) \times \mathrm{Mat}_2(\mathbb{F}_4)$, the non-commutative component arises from the Frobenius-squared action of $\mathbb{Z}_2$ on $\mathbb{F}_{16}$.

C Non-target preservation (NTP)

Standard exact-match IIA can approach test accuracy even without CRT structure (since IIA $\approx \Pr[\text{model correct}]$ for random partitions). To distinguish CRT-specific independence, we introduce **Non-Target Preservation (NTP)**: after patching component i, we check whether non-target components $j \neq i$ retain their base values.

Table 12: NTP distinguishes CRT structure from mere generalization quality.

Model	test_acc	True CRT NTP	Random NTP
Φ_7 (CRT: $\mathbb{F}_8 \times \mathbb{F}_8$), s42	0.983	**0.995**	0.159 ± 0.011
baseline$_5$ ($\mathbb{F}_{16}$, no CRT), s2048	1.000	—	0.281 ± 0.022
$\mathbb{F}_2[x]/(x^7)$ (local ring), s42	0.951	—	0.272 ± 0.007

The CRT model achieves NTP = 0.995 for the chosen decomposition, while random partitions yield NTP $\approx$ 0.16–0.28. Even the perfectly generalizing baseline$_5$ ($\mathbb{F}_{16}$, irreducible, no CRT) has much lower preservation under these controls. This supports specificity of the measured projection organization beyond task accuracy, while leaving the downstream implementation to the intermediate tests.

D Within-component algorithm analysis

Beyond component organization, we ask how model outputs agree with candidate within-component computations. Logit similarity is a functional comparison and does not uniquely identify an internal algorithm.

Definition 2 (Wedderburn circuit)**.** *A network satisfying Definition 1 further* implements a Wedderburn circuit *if, for each i, the restriction of the network's computation to the V_i channel defines a component function ϕ_i that computes the multiplication in $\mathrm{Mat}_{n_i}(D_i)$: there exists ψ_i such that $\phi_i(\pi_i(a), \pi_i(b)) = \psi_i(\pi_i(a) \cdot \pi_i(b))$.*

Correct input–output multiplication alone does not identify the implementing algorithm. We additionally report *logit cosine similarity* with a theoretically motivated circuit, while distinguishing this functional agreement from an identification of all internal operations.

Theoretical logit derivation. For each semisimple algebra $A \cong \prod_i \mathbb{F}_{q^{d_i}}$, the ideal Wedderburn circuit produces logit vectors $\ell(a, b, c) = \sum_{i=1}^{r} w_i \cdot \mathbb{1}[\pi_i(a)\pi_i(b) = \pi_i(c)]$, where, for non-zero ele-

ments, each indicator decomposes via multiplicative characters of $\mathbb{F}_{q^{d_i}}^{\times}$ (the zero element is handled separately).

Results.

Table 13: Logit cosine similarity between model output and theoretical Wedderburn circuit. Multi-seed results report mean ± std (excluding unstable seeds; see Table 16).

Algebra	Model	Fourier cos	R^2	Acc
$\mathbb{F}_2[x]/(\Phi_7) \cong \mathbb{F}_8^2$	1-layer s1024	0.735	0.556	0.999
	2-layer s789	**0.995**	**0.990**	1.000
$\mathbb{F}_2[\mathbb{Z}_7] \cong \mathbb{F}_2 \times \mathbb{F}_8^2$	1-layer s42	0.636	0.422	0.998
	2-layer (2 seeds)	**0.986** $\pm$ 0.013	**0.971** $\pm$ 0.026	1.000
$\mathbb{F}_2[\mathbb{Z}_6]$ (legacy marginals)	1-layer ($\mathbb{Z}_3$ marginal)	0.571	—	0.983
	1-layer ($\mathbb{Z}_2$ marginal)	0.119	—	0.983

The stable two-layer semisimple checkpoints achieve Fourier cosine similarity ≥ 0.976 and $R^2 \geq 0.953$, supporting agreement with the predicted logit pattern. One-layer models have lower similarity (0.636–0.735) despite high task accuracy. These comparisons do not determine a unique internal implementation; the unstable seed is retained in Table 16.

Radical-containing tasks. The historical $\mathbb{F}_2[\mathbb{Z}_6]$ logit comparisons above use the $\mathbb{Z}_2/\mathbb{Z}_3$ marginals, not the complete S/J coordinates. Low similarity to those reference logits does not identify a radical algorithm, and high raw IIA does not imply that Definition 1 has been fully verified. Nilpotent multiplication lacks the semisimple character diagonalization used in the reference calculation. The corrected coefficient interventions in Appendix K.1 establish a limited output-coordinate effect; radical WC3 remains open.

E Embedding SVD analysis

To understand why 1-layer models have high IIA but low logit similarity, we perform SVD on the embedding matrix W_E and measure η^2 (between-group variance / total variance in one-way ANOVA, groups defined by CRT component).

Table 14: Embedding SVD analysis. $\eta^2 > 0.5$ indicates a singular direction is CRT-concentrated.

Model	Top SVs w/ $\eta^2 > 0.5$	Energy in top-5	Control η^2
Φ_7 (1-layer, $\mathbb{F}_8 \times \mathbb{F}_8$)	5/5 (100%)	95%	—
$\mathbb{Z}_7$ (1-layer, $\mathbb{F}_2 \times \mathbb{F}_8^2$)	11/20 (55%)	—	—
baseline$_5$ ($\mathbb{F}_{16}$, random 2-split)	0/20	—	$\bar{\eta}^2 = 0.067$

For Φ_7, all 5 high-energy singular directions are CRT-concentrated ($\eta^2 \geq 0.618$), with SVs 0–1 dominated by factor 1 and SVs 2–4 by factor 2. For $\mathbb{Z}_7$, the three CRT factors ($\mathbb{F}_2$, $\mathbb{F}_8$, $\mathbb{F}_8$) each receive dedicated singular directions. The baseline$_5$ control ($\mathbb{F}_{16}$, irreducible) shows η^2 at the theoretical chance level ($1/15 \approx 0.067$). The singular directions are obtained without component-label supervision; the subsequent η^2 evaluation does use the labels. This is complementary evidence about input embeddings, not a substitute for intermediate causal identification or proof of a unique direct sum.

F Cross-characteristic IIA details

Table 15 provides per-seed IIA results for the three cross-characteristic algebras from Section 4.4.

G Additional logit similarity data

Seed 789 experienced training instability (model accuracy degraded to 0.912 after grokking), leading to lower logit similarity. Excluding this outlier, both remaining seeds achieve Fourier cos ≥ 0.976.

Table 15: Per-seed IIA for cross-characteristic experiments (3 of 10 seeds shown for space). Full data: $\mathbb{F}_3[\mathbb{Z}_4]$ 10/10 grok, $\mathbb{F}_5[\mathbb{Z}_3]$ 7/10 grok, $\mathbb{F}_7[\mathbb{Z}_3]$ 10/10 grok.

Algebra	Seed	Grok step	Test acc	IIA (per-comp mean)
$\mathbb{F}_3[\mathbb{Z}_4]$	42	17K	0.988	0.994
	123	27K	0.995	0.997
	789	22K	0.960	0.971
$\mathbb{F}_5[\mathbb{Z}_3]$	42	9K	0.998	0.999
	123	5K	0.976	0.984
	789	33K	0.961	0.974
$\mathbb{F}_7[\mathbb{Z}_3]$	42	45K	0.997	0.996
	123	26K	0.999	0.997
	789	62K	0.955	0.968

Table 16: Complete logit similarity results for $\mathbb{F}_2[\mathbb{Z}_7]$ 2-layer models.

Seed	CRT cos	Fourier cos	R^2	Model acc
42	0.527	**0.995**	**0.990**	1.000
123	0.708	**0.976**	**0.953**	1.000
789	0.525	0.838	0.716	0.912

H Conditional IIA and training dynamics

The original conditional IIA metric is non-target preservation among records with a correct base prediction: $\Pr[\pi_k(\hat{c}_{\text{patch}}) = \pi_k(ab)\ \forall k \neq i \mid \hat{c}_{\text{base}} = ab]$. It is distinct from the base-and-source-correct conditional target and Full CF metrics used in Appendix K. Every conditional result needs its specified denominator.

We correct the original dynamics description: the step-1K value was target accuracy over 2,000 sampled records without conditioning on base correctness. It was mislabeled conditional IIA in this appendix. Moreover, the $\mathbb{Z}_2/\mathbb{Z}_3$ marginals are not complete CRT or S/J coordinates. Appendix M.1 gives a separate exact diagnostic with corrected CRT factors and base-correct counts; Section 5.3 uses held-out output-coordinate accuracy for the equal-cardinality S/J comparison. None of these full-token diagnostics alone identifies an intermediate component circuit.

I S_5 per-seed IIA data

Table 17 provides per-seed IIA results for S_5 group multiplication, complementing the summary in Section 4.6.

Table 17: S_5: dominant Wedderburn component per seed (7 of 16 grokked seeds shown, ordered by FVE). λ-IIA measured by patching the SVD-derived subspace for each irrep.

Seed	Dominant λ	Mat_n	FVE	λ-IIA	Chance	Above chance
123	$(3, 1^2)$	$\text{Mat}_6(\mathbb{F}_7)$	0.978	**0.992**	0.50	$+0.49$
12345	$(3, 1^2)$	$\text{Mat}_6(\mathbb{F}_7)$	0.978	**0.978**	0.50	$+0.48$
42	$(2, 1^3)$	$\text{Mat}_4(\mathbb{F}_7)$	0.890	0.635	0.30	$+0.34$
2048	$(3, 1^2)$	$\text{Mat}_6(\mathbb{F}_7)$	0.718	0.703	0.50	$+0.20$
1024	$(4, 1)$	$\text{Mat}_4(\mathbb{F}_7)$	0.560	0.440	0.31	$+0.13$
7777	$(3, 1^2)$	$\text{Mat}_6(\mathbb{F}_7)$	0.520	0.526	0.50	$+0.03$
789	$(3, 1^2)$	$\text{Mat}_6(\mathbb{F}_7)$	0.382	0.530	0.50	$+0.03$

In the displayed runs, FVE ≥ 0.7 accompanies λ-IIA gaps of at least $+0.20$. Moderate-FVE examples range from $+0.03$ to $+0.13$, including an above-chance result at FVE 0.560. These descriptive comparisons show graded component-specific effects; they do not establish a sharp threshold or a quantified correlation across all components.

J Training dynamics

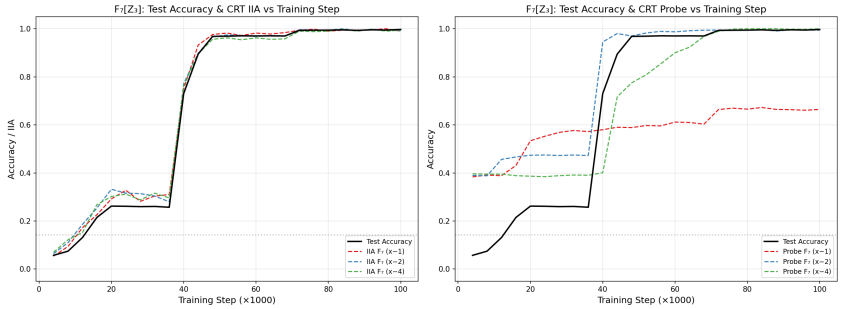


Figure 3: Training dynamics for $\mathbb{F}_7[\mathbb{Z}_3]$ (semisimple, 3 components). Raw full-swap IIA for all three $\mathbb{F}_7$ components rises with test accuracy during grokking; these curves are descriptive accuracy diagnostics.

$\mathbb{F}_7[\mathbb{Z}_3]$ **(semisimple, 3 components).** The original full-swap IIA curves rise with test accuracy around step 45K (Figure 3) and exceed 0.99 at convergence. Since full swaps can track ordinary accuracy, this co-occurrence is a descriptive training diagnostic and does not establish that internal decomposition causes grokking. The corrected $\mathbb{F}_2[\mathbb{Z}_6]$ analysis uses a different, explicitly defined output-coordinate metric (Section 5.3).

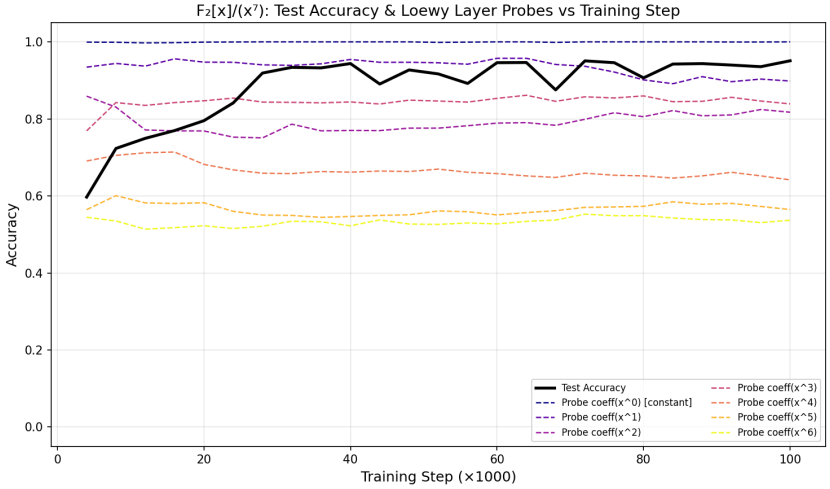


Figure 4: Training dynamics for $\mathbb{F}_2[x]/(x^7)$ (local ring, Loewy length 7). Probe accuracy generally decreases for higher Loewy powers; it is not a strictly monotone ordering.

$\mathbb{F}_2[x]/(x^7)$ **(local ring, Loewy length 7).** Loewy-layer probes show a broad accessibility trend (Figure 4): the constant term (x^0) is perfectly readable throughout training (accuracy 1.000), with accuracy generally decreasing for higher powers (x^1: 0.90, x^2: 0.82, x^3: 0.84, x^4: 0.65, x^5: 0.57, x^6: 0.54). The x^2/x^3 reversal prevents a strict-hierarchy claim; these are decoding results, not within-component causal identification.

K Intermediate component interventions and identification sensitivity

Scope and checkpoints. We examine the answer-token residual stream immediately after attention in three one-layer Transformers: $\mathbb{F}_2[\mathbb{Z}_7]$, $\mathbb{F}_2[\mathbb{Z}_6]$, and $\mathbb{F}_3[\mathbb{Z}_4]$. Each model has $d_{\text{model}} = 128$, four attention heads, and an MLP of width 512, and was trained with seed 42, a 30% training split, batch size 512, AdamW learning rate 10^{-3}, weight decay 1, and optimizer coefficients $(0.9, 0.98)$. We use the final checkpoints of these runs: the first two runs completed their $100{,}000$-update budgets;

the third stopped at the logged step 30,000. The analysis was performed on the MPS backend. These are three selected single-seed case studies, not a multi-seed estimate of intermediate alignment prevalence.

The component coordinates are

$$\mathbb{F}_2[\mathbb{Z}_7] \cong \mathbb{F}_2 \times \mathbb{F}_{8,a} \times \mathbb{F}_{8,b}, \tag{1}$$

$$\mathbb{F}_3[\mathbb{Z}_4] \cong \mathbb{F}_{3,a} \times \mathbb{F}_{3,b} \times \mathbb{F}_9, \tag{2}$$

$$\mathbb{F}_2[\mathbb{Z}_6] \cong \underbrace{\mathbb{F}_2[x]/((x+1)^2)}_{L_4} \times \underbrace{\mathbb{F}_2[x]/((x^2+x+1)^2)}_{L_{16}}. \tag{3}$$

Here L_4 and L_{16} denote local rings of cardinality 4 and 16; they are not semisimple field factors. The joint 4×16 coordinates distinguish all 64 elements. Consequently, preserving the complementary factor has its full CRT meaning. The smaller $\mathbb{Z}_2/\mathbb{Z}_3$ marginal projections do not jointly distinguish these 64 elements and are not used as a complete decomposition in this analysis.

Cross-fitted identification and intervention. For each component i, we stratify the algebra's product elements by their component label $\pi_i(c)$ and hold out approximately 20% of the complete elements. All input pairs whose product is a held-out element are excluded from fitting the probe. A linear classifier predicts $\pi_i(ab)$ from the post-attention answer-token activation $h(a, b)$. For the default setting, we use logistic regression with $C = 1$ and at most 4,000 iterations. We center the multiclass coefficient rows and take the smallest number of right singular vectors accounting for 95% of their squared-singular-value energy; for a binary classifier we retain its single coefficient direction. Writing these orthonormal directions as the columns of V_i, the same fixed projector is used for all interventions on that component:

$$h^{\text{patch}} = h(a, b) + V_i V_i^\top \big(h(a', b) - h(a, b)\big). \tag{4}$$

The base and source inputs differ only in the ith coordinate of a; b and every other coordinate are fixed. Both products ab and $a'b$ must be held-out elements, and their target coordinates must differ. The desired counterfactual is $a'b$. We sample 2,000 such records per component, with repeated records allowed, and evaluate 20 same-dimensional random orthonormal subspaces on the identical records. The split and intervention seeds are $42 + i$ (zero-based component index), and the random-subspace seed is $1000 + i$. The repeated-record sampling means that 2,000 is a protocol sample count, not a count of independent distinct interventions.

Metrics and conditioning. Target accuracy checks $\pi_i(\hat{c}) = \pi_i(a'b)$; non-target preservation (NTP) checks every $\pi_j(\hat{c}) = \pi_j(ab)$ for $j \neq i$; Full CF checks the entire counterfactual, $\hat{c} = a'b$. The unconditional versions use all 2,000 records. The conditional versions use only records for which the unpatched model predicts both ab and $a'b$ correctly; their denominator is n_{correct}. Because the target coordinate is required to change, leaving a correct base prediction unchanged gives zero conditional target accuracy. Accordingly, the random-subspace score is an intervention control, not the $1/|A_i|$ chance accuracy of an independent random label classifier. The random values below are means over 20 subspaces; they do not represent variation across model-training seeds.

Table 18: All eight components at the default logistic/95%-SVD setting. Probe is held-out decoding accuracy. All intervention metrics in this table are *unconditional*. Random target is the mean target accuracy of same-dimensional random subspaces. Local-ring labels are defined above.

Algebra	Component	Dim.	Probe	Target	Full CF	NTP	Random target
$\mathbb{F}_2[\mathbb{Z}_7]$	$\mathbb{F}_2$	1	1.0000	1.0000	0.9935	0.9935	0.000000
	$\mathbb{F}_{8,a}$	2	0.6428	0.2890	0.0785	0.3160	0.000150
	$\mathbb{F}_{8,b}$	2	0.2888	0.0665	0.0320	0.6870	0.000000
$\mathbb{F}_2[\mathbb{Z}_6]$	L_4	1	0.9565	0.3570	0.0075	0.6430	0.000000
	L_{16}	3	0.6067	0.6195	0.5640	0.8485	0.000125
$\mathbb{F}_3[\mathbb{Z}_4]$	$\mathbb{F}_{3,a}$	2	0.9815	0.8630	0.7370	0.7510	0.000000
	$\mathbb{F}_{3,b}$	1	0.6485	0.0095	0.0095	1.0000	0.011650
	$\mathbb{F}_9$	2	0.3789	0.2845	0.2140	0.7745	0.000000

The results are mixed. The $\mathbb{F}_2$ component of $\mathbb{F}_2[\mathbb{Z}_7]$ has high target transfer and non-target preservation. The $\mathbb{F}_{3,a}$ component and L_{16} exhibit appreciable but incomplete selectivity at the default

Table 19: Conditional counterparts of Table 18, restricted to records with correct unpatched base and source predictions; n_{correct} is the retained count out of $2{,}000$.

Algebra	Component	n_{correct}	Conditional target	Conditional Full CF
$\mathbb{F}_2[\mathbb{Z}_7]$	$\mathbb{F}_2$	2000	1.0000	0.9935
	$\mathbb{F}_{8,a}$	1880	0.2989	0.0835
	$\mathbb{F}_{8,b}$	1927	0.0690	0.0332
$\mathbb{F}_2[\mathbb{Z}_6]$	L_4	2000	0.3570	0.0075
	L_{16}	1849	0.6220	0.5776
$\mathbb{F}_3[\mathbb{Z}_4]$	$\mathbb{F}_{3,a}$	2000	0.8630	0.7370
	$\mathbb{F}_{3,b}$	1964	0.0000	0.0000
	$\mathbb{F}_9$	1996	0.2841	0.2134

setting. Other components remain weak. In particular, decoding $\mathbb{F}_{3,b}$ at 0.6485 does not yield a successful causal intervention: its unconditional target accuracy is 0.0095, compared with 0.01165 for the random control, and its conditional target accuracy is zero. Thus supervised decodability alone does not establish a component's causal role. Weak results are inconclusive about the existence of a different, possibly nonlinear or distributed, realization; they neither establish nor refute WA2 in general.

Complete threshold and probe-family sweep. We repeat identification with retained-energy thresholds $80\%, 90\%, 95\%$, and 99%, using either the logistic probe above or a ridge classifier with regularization $\alpha = 1$. The checkpoint, split, intervention records, and random-control protocol remain fixed. Table 20 reports every component and setting, including the dimensions selected by each rule.

Table 20: Complete probe-family and SVD-threshold sensitivity analysis. Each entry is unconditional Full CF, with subspace dimension in parentheses. All 64 settings use $2{,}000$ sampled interventions per component. These are identification-method variations on the same three trained checkpoints, not independent model runs.

Algebra	Component	Probe	80%	90%	95%	99%
$\mathbb{F}_2[\mathbb{Z}_7]$	$\mathbb{F}_2$	Logistic	0.9935 (1)	0.9935 (1)	0.9935 (1)	0.9935 (1)
	$\mathbb{F}_{8,a}$	Logistic	0.0220 (1)	0.0785 (2)	0.0785 (2)	0.1205 (4)
	$\mathbb{F}_{8,b}$	Logistic	0.0080 (1)	0.0080 (1)	0.0320 (2)	0.0320 (3)
$\mathbb{F}_2[\mathbb{Z}_6]$	L_4	Logistic	0.0075 (1)	0.0075 (1)	0.0075 (1)	0.0355 (2)
	L_{16}	Logistic	0.3215 (2)	0.5640 (3)	0.5640 (3)	0.9300 (4)
$\mathbb{F}_3[\mathbb{Z}_4]$	$\mathbb{F}_{3,a}$	Logistic	0.7370 (2)	0.7370 (2)	0.7370 (2)	0.7370 (2)
	$\mathbb{F}_{3,b}$	Logistic	0.0095 (1)	0.0095 (1)	0.0095 (1)	0.0330 (2)
	$\mathbb{F}_9$	Logistic	0.1480 (1)	0.2140 (2)	0.2140 (2)	0.4175 (4)
$\mathbb{F}_2[\mathbb{Z}_7]$	$\mathbb{F}_2$	Ridge	0.8875 (1)	0.8875 (1)	0.8875 (1)	0.8875 (1)
	$\mathbb{F}_{8,a}$	Ridge	0.0000 (1)	0.0000 (2)	0.0000 (2)	0.0000 (3)
	$\mathbb{F}_{8,b}$	Ridge	0.0000 (2)	0.0000 (2)	0.0000 (3)	0.0000 (4)
$\mathbb{F}_2[\mathbb{Z}_6]$	L_4	Ridge	0.0000 (1)	0.0000 (1)	0.0000 (1)	0.0000 (2)
	L_{16}	Ridge	0.0000 (1)	0.0000 (3)	0.0000 (4)	0.0470 (6)
$\mathbb{F}_3[\mathbb{Z}_4]$	$\mathbb{F}_{3,a}$	Ridge	0.0000 (2)	0.0000 (2)	0.0000 (2)	0.0000 (2)
	$\mathbb{F}_{3,b}$	Ridge	0.0115 (1)	0.0130 (2)	0.0130 (2)	0.0130 (2)
	$\mathbb{F}_9$	Ridge	0.0000 (1)	0.0000 (2)	0.0000 (3)	0.1500 (5)

The $\mathbb{F}_2$ result remains high across both probe families and all four thresholds. In contrast, L_{16} changes from Full CF 0.3215 to 0.9300 as the logistic threshold increases, while ridge reaches only 0.0470 at 99%. The $\mathbb{F}_{3,a}$ result is constant across logistic thresholds but vanishes with ridge. These differences restrict the interpretation to the identified, method-dependent subspaces; they do not support a uniform, probe-independent direct-sum recovery. Cross-fitting separates subspace fitting from the evaluated product elements, and same-dimensional random controls test whether arbitrary directions suffice. Neither makes the supervised identification rule independent of the proposed algebraic labels. Natural input replacement and full-state transport are reference controls, not additional evidence for a selective component subspace.

K.1 Radical output-coordinate interventions

Within $L_4 \cong \mathbb{F}_2[\epsilon]/(\epsilon^2)$, write $a = a_0 + a_1\epsilon$ and $b = b_0 + b_1\epsilon$. The output coefficients satisfy

$$c_0 = a_0 b_0, \qquad c_1 = a_0 b_1 + a_1 b_0 \quad \text{in } \mathbb{F}_2. \tag{5}$$

We test whether a low-dimensional subspace mediates changes in c_1, while fixing c_0 and the complete complementary L_{16} output. This exploratory analysis uses the same $\mathbb{F}_2[\mathbb{Z}_6]$ checkpoint, whose accuracy over all $4{,}096$ input pairs is 0.976074. Unlike the product-element holdout above, this experiment randomly partitions input pairs into $2{,}867$ fitting pairs and $1{,}229$ held-out pairs (seed 42). On the fitting set, we collect $10{,}000$ matched activation differences and retain their leading singular directions at 95% energy. Evaluation uses $2{,}000$ held-out matched records and 20 same-dimensional random subspaces. The resulting basis is fixed across evaluated pairs at each site.

We also form matched pairs by toggling either theoretical term $t_{01} = a_0 b_1$ or $t_{10} = a_1 b_0$, while holding the other term, $a_0 b_0$, and the L_{16} output fixed. In both cases the evaluated target remains the nilpotent output coefficient c_1; the metric is not a measurement of a separately represented internal multiplication term. Table 21 reports these tests at all four recorded sites.

Table 21: Corrected radical-coordinate analysis with the full 4×16 CRT decomposition. Target and Full CF are conditional on correct unpatched base and source predictions. The conditional counts are $1{,}878$, $1{,}889$, and $1{,}895$ for t_{01}, t_{10}, and c_1 pairing, respectively, out of $2{,}000$ records each; they are constant across sites. Random conditional target is the mean over 20 same-dimensional random subspaces. All target metrics evaluate c_1.

Site	Pairing target	Dim.	Cond. target	Cond. Full CF	Random cond. target
Post-attention	t_{01}	4	0.9840	0.9239	0.000186
	t_{10}	4	0.9889	0.9121	0.000926
	c_1	4	0.9884	0.9182	0.000844
MLP pre-activation	t_{01}	1	0.9010	0.7471	0.000000
	t_{10}	1	0.9068	0.7263	0.000000
	c_1	1	0.9050	0.7420	0.000000
MLP GELU hidden	t_{01}	3	0.9286	0.9063	0.000000
	t_{10}	3	0.9344	0.9079	0.000053
	c_1	3	0.9330	0.9140	0.000000
Post-MLP	t_{01}	2	0.9366	0.9196	0.000000
	t_{10}	2	0.9439	0.9222	0.000185
	c_1	2	0.9340	0.9193	0.000026

For the post-attention c_1 test specifically, unconditional target accuracy is 0.9770 and unconditional Full CF is 0.9040; preservation of c_0 and L_{16} is 0.9965 and 0.9225, respectively. On the $1{,}895$ records with correct base and source predictions, target accuracy is 0.9884 and Full CF is 0.9182. The corresponding random target means are 0.012975 unconditional and 0.000844 conditional. This distinction is necessary when comparing the high conditional target score with preservation metrics computed over all records.

These results support a low-dimensional mediator of the nilpotent *output coordinate* in this checkpoint. They do not identify separate internal representations of $a_0 b_1$ and $a_1 b_0$, their order of computation, or a unique truncated-convolution algorithm. In particular, success when pairs are grouped by a theoretical term does not by itself show that the network explicitly computes that term. The learned within-component algorithm, and hence the stronger radical WC3 interpretation, remain unresolved.

L Additional capacity and optimization controls for $\mathbb{F}_2[D_5]$

The original $0/12$ result concerns four configurations ($d = 128, L = 1$; $d = 128, L = 2$; $d = 256, L = 2$; and $d = 512, L = 1$), each trained with seeds 42, 123, and 789 at learning rate 10^{-3} for 100K steps. It is not a claim that $\mathbb{F}_2[D_5]$ cannot grok. We subsequently tested a complete seed-42 grid at learning rate 10^{-4} and a 300K-step horizon, followed by two additional seeds for $d = 1024, L = 2$.

The new grid uses $d \in \{128, 256, 512, 1024\}$, $L \in \{1, 2, 4\}$, $d_{\mathrm{MLP}} = 4d$, and $d/64$ attention heads, without LayerNorm. Training uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.98)$, weight decay 1.0, and batch size 512. A seed-specific permutation of the 1024^2 ordered input pairs supplies 314,572 training pairs (30% of the table, rounded down) and 50,000 disjoint test pairs. Grokking time is the first recorded test accuracy at least 0.95, evaluated at 1K-step intervals; this definition does not impose a sustained threshold. The entries called final accuracy below are the last logged evaluations of the 300K-step runs. The new grid is an expanded protocol: its horizon, learning rate, and, for the narrowest model, head count differ from the original settings. It should not be read as a one-variable comparison for every architecture.

Table 22: Complete $\mathbb{F}_2[D_5]$ width–depth grid at learning rate 10^{-4}, seed 42, and a 300K-step horizon. A dash means that the test accuracy never reached 0.95 within the recorded evaluations.

d	L	Parameters (M)	Grok step	Final train	Final test
128	1	0.459	—	0.126184	0.118300
128	2	0.656	—	0.185519	0.160620
128	4	1.049	—	0.162529	0.145360
256	1	1.312	—	0.265701	0.225620
256	2	2.098	—	0.313706	0.266760
256	4	3.671	—	0.349112	0.286720
512	1	4.196	—	0.377551	0.273880
512	2	7.342	—	0.339042	0.288260
512	4	13.634	112K	0.999619	0.999480
1024	1	14.684	34K	1.000000	1.000000
1024	2	27.267	25K	0.999409	0.999400
1024	4	52.433	13K	0.998798	0.998620

Table 23: Three seeds for $\mathbb{F}_2[D_5]$, $d = 1024, L = 2$, learning rate 10^{-4}, and 300K training steps. Seed 42 is the same run as in Table 22, not an additional replication.

Seed	Grok step	Final train	Final test
42	25K	0.999409	0.999400
123	34K	0.998620	0.998320
789	31K	0.999758	0.999800

Under this common low-learning-rate protocol, all tested $d = 128$ and $d = 256$ models remain below threshold, $d = 512$ succeeds only at $L = 4$, and $d = 1024$ succeeds at all three tested depths. The $d = 1024, L = 2$ configuration succeeds for all three tested seeds, whereas the remaining grid cells have only one seed. The delayed 112K transition of $d = 512, L = 4$ also shows that a 100K horizon can miss a successful trajectory. Together these observations support an empirical interaction among capacity, optimization, and training duration. They do not isolate non-commutativity as the causal source of a universal capacity threshold. These new runs establish successful generalization; they have not been analyzed for component-level alignment and therefore add no direct alignment claim.

M Corrected output-coordinate dynamics in $\mathbb{F}_2[\mathbb{Z}_6]$

Complete semisimple and radical coordinates. Using $\mathbb{Z}_6 \cong \mathbb{Z}_2 \times \mathbb{Z}_3$, write $A = \mathbb{F}_2[\mathbb{Z}_6] \cong S[g]/(g^2 - 1)$, where $S = \mathbb{F}_2[\mathbb{Z}_3]$ and g generates the $\mathbb{Z}_2$ factor. Set $t = g + 1$, so that $t^2 = 0$. Every element has a unique expression

$$a = r_0 + gr_1 = (r_0 + r_1) + tr_1 = s + tj, \qquad s, j \in S.$$

Here S is semisimple, $J = tS$ is the Jacobson radical, and both the semisimple coordinate s and the radical coordinate j take eight values (three bits). The map $a \mapsto (s, j)$ is a bijection onto $S \times S$ as a set and vector space. It is not a direct-product decomposition of rings:

$$(s + tj)(s' + tj') = ss' + t(sj' + js'). \tag{6}$$

The radical coordinate records the coefficient of t; the multiplication on J itself is zero. An exhaustive check verified all 64 distinct joint coordinates and Equation 6 for all 4,096 products.

Protocol and what is measured. We train three one-layer Transformers with $d = 128$, four heads, and $d_{\mathrm{MLP}} = 512$, without LayerNorm, using seeds 42, 123, and 789. Each run uses AdamW, learning rate 10^{-3}, weight decay 1.0, $(\beta_1, \beta_2) = (0.9, 0.98)$, batch size 512, and 100K steps. For each seed, the original random pair split supplies 1,228 training pairs and all 2,868 remaining pairs as the held-out test set. Checkpoints are evaluated every 1K steps, including the 100K final model.

For a test input (a, b), let $\widehat{c}$ be the model's predicted product and $c = ab$ the true product. We compare $s(\widehat{c})$ with $s(c)$ and $j(\widehat{c})$ with $j(c)$ directly. No learned probe, hidden activation, or intervention enters this measurement. It measures output-coordinate learning on held-out operand pairs, not the identification of independent internal causal subspaces. Both targets have eight classes, but their class frequencies differ. The primary metric is therefore balanced accuracy, the unweighted mean recall over the eight true output-coordinate classes. The test-set majority baselines for s are 0.326, 0.324, and 0.327 for the three seeds, versus 0.185, 0.191, and 0.187 for j. We define sustained emergence as the first of three consecutive 1K-spaced checkpoints whose coordinate metric is at least 0.8. This is a three-checkpoint criterion, not a requirement that the metric remain above threshold forever. Grokking separately means first full-product test accuracy at least 0.95.

Table 24: All three $\mathbb{F}_2[\mathbb{Z}_6]$ trajectories. Emergence times use three consecutive checkpoints at or above 0.8. The gap is radical minus semisimple emergence time under the primary balanced-accuracy metric. Only seed 42 reaches the full-task grokking threshold by 100K.

		Balanced accuracy			Raw accuracy	
Seed	Grok step	s	j	Gap	s	j
42	24K	9K	12K	3K	8K	11K
123	—	11K	25K	14K	10K	24K
789	—	7K	8K	1K	6K	8K

Table 25: Held-out accuracy at the actual 100K final checkpoint of each trajectory. BA denotes balanced accuracy. These are direct evaluations of the final model, not the last periodic training-log entry (99K).

Seed	Full product	s accuracy	s BA	j accuracy	j BA
42	0.986402	0.998954	0.998794	0.986750	0.988374
123	0.887029	0.995816	0.990990	0.888075	0.880207
789	0.936541	0.998954	0.999074	0.937238	0.933720

The semisimple output coordinate reaches the threshold earlier in all three trajectories, including the only grokked seed. Raw accuracy gives the same ordering. This removes the unequal-cardinality explanation of the earlier comparison, while balanced accuracy addresses the unequal label frequencies. The result is consistent with the triangular dependency in Equation 6: the semisimple output depends only on semisimple inputs, whereas the radical output depends on both coordinates. It does not establish that this dependency causes the observed timing, or that a universal ordering holds across grokked seeds and algebras.

M.1 CRT-factor and metric-label audit

The earlier $\mathbb{Z}_2/\mathbb{Z}_3$ quotient marginals jointly take only 16 values and do not invertibly encode the 64 elements of $\mathbb{F}_2[\mathbb{Z}_6]$. In particular, the four-valued $\mathbb{Z}_2$ marginal is not the full radical coordinate. The correct CRT factors are

$$\mathbb{F}_2[x]/(x^6 + 1) \cong \mathbb{F}_2[x]/((x + 1)^2) \times \mathbb{F}_2[x]/((x^2 + x + 1)^2),$$

of sizes 4 and 16. Both factors are local rings with nonzero radical; they are not a semisimple factor paired with a radical factor. Their joint projection is bijective, and multiplication compatibility was checked exhaustively on all 4,096 products.

We also correct a metric label: the earlier step-1K value was raw target-coordinate accuracy over 2,000 sampled interventions, not conditional IIA. For a separate audit using the correct CRT factors, we enumerate every triple (a, a', b) with the non-target input coordinate fixed, the target input

coordinate changed, and the target product coordinate changed. There are 8,192 triples for the four-element target and 55,296 for the 16-element target. These enumerations range over the full multiplication table, including training pairs; unlike the held-out coordinate analysis above, they are descriptive full-table diagnostics. Full-token embedding replacement is equivalent to evaluating (a', b), so all metrics are computed exactly from the corresponding prediction table.

Raw target accuracy checks the target coordinate of the counterfactual prediction; Full CF checks the complete counterfactual product; NTP checks the non-target coordinate against the base product. Conditional NTP restricts the last check to triples for which the base prediction is correct. Table 26 reports all three seeds at 1K.

Table 26: Correct-CRT diagnostics at step 1K. n_{base} is the number of base-correct intervention triples, the denominator of conditional NTP. Counts are intervention records, not independent test examples.

Seed	Target size	Raw target	Full CF	NTP	Cond. NTP	n_{base}
42	4	0.830	0.429	0.484	0.582	3,513
42	16	0.445	0.413	0.881	0.898	22,860
123	4	0.848	0.432	0.481	0.597	3,542
123	16	0.433	0.407	0.895	0.903	22,488
789	4	0.715	0.271	0.349	0.639	2,220
789	16	0.304	0.257	0.793	0.859	14,238

Under the same sustained 0.8 criterion, raw target accuracy reaches threshold at 1K/1K/2K for the four-element factor and 9K/18K/6K for the 16-element factor (seeds 42/123/789). Full CF instead reaches threshold at 11K/24K/9K when targeting the four-element factor and 10K/22K/7K when targeting the 16-element factor. Thus the ordering depends on whether success requires only the targeted output coordinate or the whole counterfactual product. Conditional NTP can already be high at initialization because of concentrated prediction marginals, so we do not use it alone to define emergence. The semisimple-before-radical finding in Table 24 comes from the complete, equal-sized (s, j) coordinates and held-out predictions, not from relabeling these unequal CRT factors.

N Embedding intervention protocol

Full-swap diagnostic. The original protocol patches the embedding at `hook_resid_pre_0`. For target projection π_i, inputs (a, a') match on all measured non-target projections, share a second operand b, and differ on the target. We replace the entire position-0 activation with its source value. For genuine CRT factors, the counterfactual label $a'b$ differs from ab only in the target factor. This specifies an algebraic input counterfactual, not an isolated neural subspace. The historical $\mathbb{F}_2[\mathbb{Z}_6]$ marginal projections are distinguished from CRT factors in Appendix M.1. Since the entire activation is replaced, raw IIA can approach ordinary test accuracy without structured decomposition and is a sanity check.

Preservation metrics. Error-case preservation uses records where the base model's full-product prediction is wrong and compares each non-target coordinate of the patched prediction with that of the base prediction. The preserved coordinate may itself be correct; the filter does not require a component-level error. The reference rate is $1/|\operatorname{im} \pi_k|$. Above-reference preservation describes structured behavior on base-error examples without establishing an independent error process for every factor. The original conditional IIA instead measures non-target preservation on base-correct records. Its denominator varies with accuracy and must be reported, particularly early in training. The earlier step-1K dynamics value was unconditional target accuracy (Appendix H).

Embedding subspace identification. For Table 2, we collect position-0 embeddings of all algebra elements and fit a logistic regression probe to predict $\pi_i(a)$. We compute the SVD of its coefficient matrix $W \in \mathbb{R}^{|\operatorname{im} \pi_i| \times d_{\text{model}}}$ and retain the top right singular vectors capturing 95% of its squared-singular-value energy. They form an orthonormal basis V of effective dimension d_{eff}. Patching is restricted to that subspace:

$$h_{\text{patched}} = h_{\text{base}} + VV^{\top}(h_{\text{source}} - h_{\text{base}}).$$

The later intermediate analysis uses disjoint product-element partitions and an explicit probe-family and threshold sweep, as detailed in Appendix K.

O Additional related work

Mechanistic interpretability of grokking. Nanda et al. [2023] identified DFT circuits in Transformers trained on $\mathbb{Z}_{113}$ addition. Chughtai et al. [2023] extended this to group operations, showing GCR-type circuits for several groups including S_5. Our S_5 experiments use the same group and real-valued neural representations, using its semisimple decomposition over $\mathbb{F}_7$ as an algebraic reference and adding component-specific intervention tests. Zhong et al. [2023] identified "clock" and "pizza" algorithms for modular arithmetic. Stander et al. [2024] analyzed coset-based algorithms for group multiplication. Our framework connects DFT/GCR/CRT descriptions and extends their empirical investigation to finite rings with radical; it does not claim to identify every clock, pizza, or coset algorithm as the same circuit.

Grokking theory. Liu et al. [2022] proposed representation learning as the mechanism behind grokking. Morwani et al. [2024] proved that Fourier features emerge as max-margin solutions under weight decay for cyclic groups. Gromov [2023] characterized grokking in quadratic networks. Tikeng Notsawo et al. [2024] analyzed loss-landscape predictors of grokking. Our work extends the empirical scope to ring algebras with radical and reports configuration-dependent grokking in selected non-commutative tasks, without deriving their training dynamics.

Causal interpretability. Geiger et al. [2021, 2024] developed IIA as a causal method for testing whether neural representations implement posited computational structures. Wu et al. [2023] applied causal methods at scale. We adapt IIA to algebraic decompositions and introduce NTP to distinguish structural independence from high accuracy.

NeurIPS Paper Checklist

1. **Claims**

   Answer: [Yes]

   Justification: Abstract and Introduction distinguish the formal framework from empirical evidence, and report mixed intermediate identification. The Limitations section separates method dependence, task coverage, scale, and correlational claims.

2. **Limitations**

   Answer: [Yes]

   Justification: The Limitations section explicitly discusses classical-algebra novelty, input versus intermediate alignment, probe dependence, non-commutative coverage, radical algorithms, confounded associations, and release coverage.

3. **Theory assumptions and proofs**

   Answer: [Yes]

   Justification: Propositions 1–3 state full assumptions. Proof sketches in Appendix A.

4. **Experimental result reproducibility**

   Answer: [Yes]

   Justification: Section 4.1 specifies model architecture, optimizer, learning rate, weight decay, batch size, and training steps. All algebra constructions are deterministic from their mathematical definitions.

5. **Open access to data and code**

   Answer: [No]

   Justification: A public self-contained package provides the three representative training tasks and intermediate component-subspace analysis, but it does not cover every historical result. Tables are deterministically generated from the algebraic definitions; package scope is stated in the experimental setup.

6. **Experimental setting/details**

   Answer: [Yes]

   Justification: Section 4.1 and table captions specify all hyperparameters, seed counts, and IIA protocol details.

7. **Experiment statistical significance**

   Answer: [Yes]

   Justification: Multi-seed outcomes and denominators are reported explicitly. The intermediate analysis uses one trained seed per algebra; random-subspace variation is not represented as training-seed uncertainty. The full identification sensitivity sweep and weak components are included.

8. **Experiments compute resources**

   Answer: [Yes]

   Justification: Original experiments used a single A100 80GB GPU per run, with a previously estimated 200 GPU-hours. Added intermediate and coordinate analyses used MPS; additional D5 runs used GPUs and up to 300K updates per run. The 200-hour estimate excludes these added runs; aggregate additional wall-clock usage was not logged centrally.

9. **Code of ethics**

   Answer: [Yes]

   Justification: This is foundational research on mathematical structure in neural networks with no direct negative societal applications.

10. **Broader impacts**

    Answer: [N/A]

Justification: This is theoretical/empirical research on mechanistic interpretability of small Transformers learning algebraic operations. No direct societal impact.

11. **Safeguards**

    Answer: [N/A]

    Justification: No models or datasets with risk of misuse are released.

12. **Licenses for existing assets**

    Answer: [N/A]

    Justification: No pre-existing datasets or code packages are used. All data is generated from mathematical definitions.

13. **New assets**

    Answer: [Yes]

    Justification: The representative code package documents algebra generators, configurations, seeds, metric definitions, and runtime dependencies; it distributes neither checkpoints nor precomputed results.

14. **Crowdsourcing and research with human subjects**

    Answer: [N/A]

    Justification: No human subjects or crowdsourcing involved.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Answer: [N/A]

    Justification: No human subjects research.

16. **Declaration of LLM usage**

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core research methodology.