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]

Abstract

Prior work has shown that grokked Transformers implement discrete Fourier transforms or group character representations for modular arithmetic. We unify these findings under the Wedderburn–Artin decomposition: using interchange intervention accuracy (IIA), a causal method stronger than ablation, we show that grokked models decompose computation along the Wedderburn components of the target algebra. Across 11 commutative algebras over $\mathbb{F}_p$ ($p = 2, 3, 5, 7$)—including 6 with nontrivial Jacobson radical—all grokked models achieve raw IIA ≥ 0.97. When models predict incorrectly, errors remain compartmentalized by component ($3.8\times$ chance), confirming that independence is structural, not a byproduct of accuracy. Training dynamics reveal that Wedderburn alignment emerges synchronously with grokking. For non-commutative groups, models exhibit *selective* Wedderburn alignment; non-commutativity raises a capacity threshold that larger models overcome.

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. Moreover, all analyses are restricted to *semisimple* algebras (groups with $\mathrm{char}(k) \nmid |G|$), leaving open the question of what happens when the algebra contains a non-trivial Jacobson radical or is non-commutative.

Contributions. We make three contributions:

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).
2. **Causal evidence via IIA.** Using IIA (Geiger et al., 2024), we verify Wedderburn alignment across 11 commutative algebras (5 semisimple, 6 radical-containing; char 2/3/5/7), 1 non-commutative algebra ($\mathbb{F}_2[S_3]$), and 1 non-commutative group (S_5, via group multiplication). A four-level baseline hierarchy rules out trivial explanations, and error compartmentalization ($3.8\times$ chance) confirms that component independence is structural. Training dynamics show that IIA rises synchronously with grokking.
3. **Grokking determinants.** A sweep across 12 algebras and multiple model capacities reveals that non-commutativity raises the capacity threshold for grokking rather than constituting an absolute barrier: $\mathbb{F}_2[S_3]$ (non-commutative, rf = 17%) groks at 4/10 seeds with d_model=128 (1-layer), 6/10 with d256 (1-layer), and 7/10 with d256 (2-layer).

2. Background

2.1. 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

[1]Microsoft Research Asia [2]Qiuzhen College, Tsinghua University [3]Department of Psychology and Cognitive Science, Tsinghua University. Correspondence to: Shurui Zheng <zhengsr23@mails.tsinghua.edu.cn>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

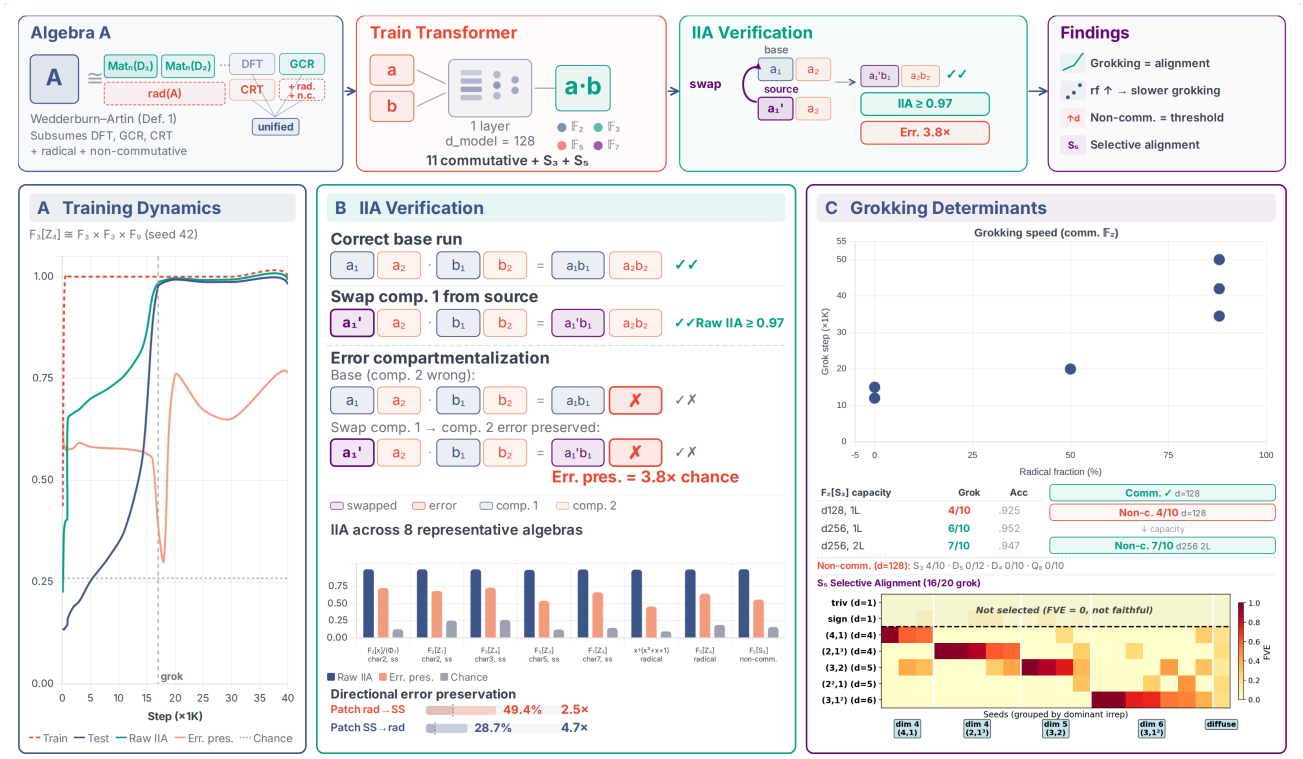

Figure 1. Top: Pipeline overview. Wedderburn-aligned representations (Def. 1) unify DFT, GCR, and CRT circuits as special cases of the Wedderburn–Artin decomposition $A/\mathrm{rad}(A) \cong \prod \mathrm{Mat}_{n_i}(D_i)$. A one-layer Transformer (d_{model}=128) is trained on algebra multiplication across 11 commutative algebras ($\mathbb{F}_p$, p=2,3,5,7), 1 non-commutative algebra ($\mathbb{F}_2[S_3]$), and S_5 group multiplication. **(A)** Training dynamics for $\mathbb{F}_3[\mathbb{Z}_4]$ (seed 42): test accuracy and IIA jump simultaneously at grokking ($\sim$17K steps). **(B)** IIA verification: component swap yields correct counterfactual; error compartmentalization (3.8$\times$ chance) confirms structural independence. Directional error preservation shows both directions well above chance. **(C)** Grokking determinants: radical fraction predicts grokking speed (Spearman $\rho = 0.35$); $\mathbb{F}_2[S_3]$ capacity sweep (4/10 $\rightarrow$ 7/10) shows non-commutativity raises a threshold; S_5 selective alignment heatmap shows each seed concentrates on 1–2 faithful irreps.

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).

2.2. 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, raw IIA can approach test accuracy without structural decomposition; we address this with **error compartmentalization** (Section 4.3).

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, $\pi_{\mathrm{rad}} : A \rightarrow \mathrm{rad}(A)$ the radical projection, and $\pi_{\mathrm{ss}} : A \twoheadrightarrow S$ the semisimple projection.*

A network $f_\theta : A \times A \rightarrow A$ with embedding $E : A \rightarrow \mathbb{R}^d$ 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 $\pi_i(b)$. Equivalently: for any $a, a', b, b' \in A$, if $\pi_i(a) = \pi_i(a')$ and $\pi_i(b) = \pi_i(b')$, then replacing the V_i-subspace activations from the (a, b) forward pass with those from (a', b') yields the correct counterfactual output.*

Remark 1 (Radical coupling)**.** *For non-semisimple A, the radical component of ab depends on semisimple components: $\pi_{\mathrm{rad}}(ab) = \pi_{\mathrm{rad}}(a)\pi_{\mathrm{ss}}(b) + \pi_{\mathrm{ss}}(a)\pi_{\mathrm{rad}}(b) + \pi_{\mathrm{rad}}(a)\pi_{\mathrm{rad}}(b)$. Despite this coupling, error compartmentalization analysis (Section 4.4) confirms that the coupling does not destroy component independence: when the base model errs, non-target components still preserve the base's wrong prediction at $3.8\times$ chance.*

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).

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 whether grokked Transformers learn Wedderburn-aligned representations (Definition 1) by measuring IIA across multiple algebras, characteristics, and training phases.

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.

IIA protocol. We perform activation patching at the embedding layer (hook_resid_pre_0). For each target Wedderburn component π_i, we select pairs (a, a') with $\pi_j(a) = \pi_j(a')$ for all $j \neq i$ but $\pi_i(a) \neq \pi_i(a')$, share a common second operand b, and replace the full position-0 activation from the source forward pass (on a') into the base forward pass (on a). Since non-target components are matched, this is equivalent to intervening only on the π_i channel: the counterfactual label is $a' \cdot b$, which differs from $a \cdot b$ only in the π_i component. We check whether the patched output matches this counterfactual. Because the intervention replaces the *entire* position-0 activation, raw IIA can approach test accuracy even without structured decomposition (Section 4.3). We therefore report two additional metrics that provide genuine discriminative power: (1) *error-case preservation*: when the base model predicts incorrectly, we check whether non-target components π_k ($k \neq i$) of the patched output preserve the base's *wrong* prediction (rather than the correct label). A preservation rate significantly above chance ($1/|\mathrm{im}\,\pi_k|$) indicates that errors are compartmentalized by component—the network "errs independently" per Wedderburn factor. (2) *Conditional IIA*: the non-target component preservation rate restricted to base-correct pairs—the fraction of interventions where $\pi_k(\text{patched}) = \pi_k(\text{base label})$ for $k \neq i$, given that the base model predicts correctly. This metric is most informative for tracking the *emergence* of component independence during training (Section 5.3), as it decouples structural alignment from overall accuracy.

Subspace identification. For the subspace-specific IIA reported in the baseline hierarchy (Table 2), we identify each component subspace $V_i \subseteq \mathbb{R}^{d_{\text{model}}}$ via probe SVD: (1) collect position-0 embeddings for all algebra elements; (2) train a logistic regression probe to predict $\pi_i(a)$; (3) compute the SVD of the probe weight matrix $W \in \mathbb{R}^{|\mathrm{im}\,\pi_i| \times d_{\text{model}}}$ and retain the top right singular vectors capturing 95% of variance. This yields an orthonormal basis of effective dimension d_{eff}. Patching is then restricted to this subspace: $h_{\mathrm{patched}} = h_{\mathrm{base}} + VV^\top(h_{\mathrm{source}} - h_{\mathrm{base}})$.

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

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
Random init	0.014	0.016	0.020
Memorized (step 500)	0.997	0.098	0.564
Grokked (step 60K)	1.000	1.000	0.998
Random subspace (same dim)	1.000	1.000	0.129 ± 0.010

Table 3. IIA for semisimple algebras. "Err. pres.": preservation rate of base's *wrong* prediction on non-target components when base is incorrect (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	$\mathbb{F}_7 \times \mathbb{F}_7 \times \mathbb{F}_7$

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 achieves IIA = 0.998, vastly exceeding both the memorized model (0.564) and the random-subspace control (0.129). The random-subspace baseline is particularly informative: even though the model achieves perfect accuracy, patching a random subspace of the same dimensionality as the CRT subspace yields IIA near chance—demonstrating that high IIA is specific to the algebraically meaningful subspace, not a byproduct of high test accuracy. As an additional negative control, $\mathbb{F}_2[x]/(x^7)$ (a local ring with no CRT decomposition) yields random-subspace IIA of 0.029 ± 0.006, confirming chance-level performance when no algebraic decomposition exists. We replicate this control on two additional algebras: $\mathbb{F}_2[\mathbb{Z}_7]$ (semisimple, 3 components) yields CRT-subspace IIA $=$ 0.992 vs. random $= 0.245 \pm 0.008$; $\mathbb{F}_2[x]/(x^4(x^3+x+1))$ (radical-containing) yields CRT-subspace IIA $= 0.801$ vs. random $= 0.129$. The lower CRT-subspace IIA for the radical-containing algebra (0.801 vs. full IIA 0.983) suggests that radical components are partially encoded nonlinearly, outside the linear subspace identified by probe SVD; nonetheless, the CRT-vs-random gap remains large ($6.2\times$). A complementary non-target preservation (NTP) metric, which checks whether non-patched components retain their base values, provides equivalent verification (Appendix C).

4.3. Multi-Algebra IIA: Semisimple Case

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

Key finding: Across all five semisimple algebras, spanning four field characteristics and component counts from 2 to 3, the mean IIA exceeds 0.98. This demonstrates that the Wedderburn decomposition governs the model's internal computation regardless of characteristic.

Modular negative control. $\mathbb{F}_3[\mathbb{Z}_3]$ (27 elements, char 3 divides $|\mathbb{Z}_3| = 3$) is a local ring with rad-fraction 2/3. It fails to grok (0/10 seeds, test accuracy ≤ 0.40), confirming that semisimplicity—not characteristic alone—enables the structured computation. This contrasts sharply with $\mathbb{F}_3[\mathbb{Z}_4]$ (same characteristic, semisimple, 10/10 grok).

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 rule this out via *error compartmentalization*: when the base model predicts incorrectly, do non-target components still preserve the (wrong) base prediction? The "Err. pres." columns in Tables 3 and 4 show that error preservation exceeds chance by $3.8\times$ on average across all tested algebras (11 algebras: 5 semisimple, 6 radical-containing). This confirms that component independence is a structural property of the network's computation, not a consequence of high accuracy.

4.4. 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).

Directional error analysis (polynomial quotient rings). We split error preservation by CRT factor type—local (radical-containing) vs. field (semisimple)—for the 4 polynomial quotient rings (Table 5).

Both directions show error preservation well above chance, confirming that the CRT factors operate as independent computational channels regardless of which factor is intervened upon.

4.5. Non-Commutative Groups

For $\mathbb{F}_2[S_3]$ (non-commutative, modular, 64 elements), grokked models at d_model=256 show complete block separation: the trivial block ($\mathbb{F}_2$, dim 3) and standard block ($\mathrm{Mat}_2(\mathbb{F}_2)$, dim 12) achieve cross-block preservation ≥ 0.999 (Table 6).

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

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.

Explanation: faithfulness + 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. Weight decay therefore penalizes redundant components, driving eff_num toward 1.0—not because the model lacks capacity, but because the task requires only one component. This is confirmed by capacity scaling (Table 7): eff_num *decreases* from 1.85 to 1.10 as d_model increases from 128 to 512. At d_model $= 512$, all grokked seeds use exactly one irrep despite having $512 \gg 120$ dimensions available.

The categorical exclusion of 1-dimensional irreps follows from the same principle: they are not faithful (trivial and sign have kernels S_5 and A_5 respectively).

Per-component IIA correlates with FVE: λ-IIA exceeds chance only when FVE ≥ 0.7. Full per-seed IIA data is in Appendix I.

5. Experiment 2: What Determines Grokking?

5.1. What Algebraic Properties Predict Grokking?

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

Radical fraction is a natural summary of algebraic complexity—the proportion of dimensions occupied by nilpotent structure. Across our data, it correlates with grokking difficulty (Spearman $\rho = 0.35$, $p = 0.001$, $n = 82$ grokked seeds): semisimple algebras (rf $= 0\%$) grok fastest with the highest success rates, while high-rf algebras grok slower and less reliably. Algebra size $|A|$ also plays a role ($|A| \geq 1024$ never groks at any tested capacity), but among algebras that do grok, rf is the more salient predictor. A structural limitation of group algebras is that rf takes only discrete values ($0, 1/2, 7/8$ for $\mathbb{F}_2$; $2/3$ for $\mathbb{F}_3$), so a fine-grained rf sweep requires polynomial quotient rings where rf can be varied continuously at fixed $|A|$.

5.2. Non-Commutativity Raises a Capacity Threshold

In Table 8, non-commutative algebras show reduced grok rates compared to commutative counterparts—even $\mathbb{F}_2[D_5]$ (rf $= 10\%$, 0/12 across 4 configurations) fails completely—but non-commutativity is not an absolute barrier: $\mathbb{F}_2[S_3]$ (rf $= 17\%$) groks at 4/10 seeds at d_model $= 128$, and increasing both width and depth improves to 7/10 (Table 9).

Non-commutativity raises the capacity threshold: increasing d_model from 128 to 256 improves grok rate from 4/10 to 6/10, and adding a second layer further improves to 7/10 while substantially reducing median grok step (14.1K $\rightarrow$ 5.4K). Combined with S_5 group multiplication (Section 4.5), where 16/20 seeds grok at d_model $= 128$, the picture is that semisimplicity—not commutativity—is the key algebraic prerequisite. (Note: S_3 and S_4 as group multiplication tasks fail to grok due to insufficient training data—5 and 173 pairs respectively—not algebraic structure; S_5 with 4320 pairs suffices.)

5.3. Training Dynamics: Radical Emerges Before Semisimple

For $\mathbb{F}_2[\mathbb{Z}_6]$ (seed 42), we tracked IIA and probe accuracy across 34 checkpoints during training (Table 10).

Surprisingly, the radical-containing $\mathbb{Z}_2$ component emerges

Table 4. IIA for algebras with nontrivial Jacobson radical. Values are mean across grokked seeds. "Err." = error preservation rate (chance in parentheses).

Algebra	$\lvert A\rvert$	char	rf	Raw Rad	Raw SS	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)

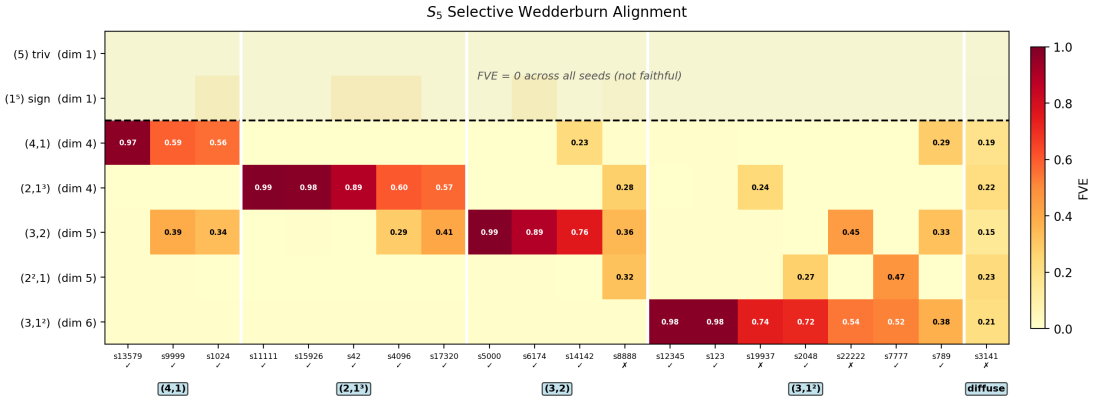

Figure 2. FVE across 20 seeds of S_5 group multiplication, grouped by dominant irrep. *Above dashed line:* the two dim=1 representations (trivial and sign) have FVE $= 0$ across all seeds. *Below:* among dim$\geq$4 irreps, each grokked seed (✓) concentrates on a different 1–2 components (block-diagonal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

Table 5. Directional error preservation for polynomial quotient rings. "Patch $\rightarrow$ Preserve": we patch the first type's activations and measure whether the second type's (wrong) prediction is preserved.

Direction	Err. pres.	Chance	Ratio
Rad $\rightarrow$ SS	0.494	0.195	2.5$\times$
SS $\rightarrow$ Rad	0.287	0.061	4.7$\times$

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 IIA	Other pres.
Trivial ($\mathbb{F}_2$)	3	0.959	0.999
Standard ($\mathrm{Mat}_2(\mathbb{F}_2)$)	12	0.999	1.000

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

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

first (IIA > 0.8 at step 1K), five times earlier than the semisimple $\mathbb{Z}_3$ component (step 6K). This is likely a complexity effect rather than a radical-vs-semisimple distinction: $\mathbb{F}_2[\mathbb{Z}_2]$ has only 4 elements (16-entry multiplication table) while $\mathbb{F}_2[\mathbb{Z}_3]$ has 8 elements (64-entry table). The $\mathbb{Z}_3$ component's emergence (steps 6K–19K) coincides precisely with grokking, suggesting that the formation of the semisimple Fourier circuit *is* the grokking transition.

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. As shown in Section 4.5, selective Wedderburn alignment in S_5 group multiplication is explained by the faithfulness of $\dim \geq 4$ irreps combined with weight decay's preference for sparse solutions. An open question is whether non-commutative *algebra* multiplication (where all components are needed) exhibits complete alignment with sufficient capacity; our $\mathbb{F}_2[S_3]$ results (Section 4.5) suggest yes, but the sample size is limited.

Table 8. Grokking results across algebras. Default: d_model=128, 1-layer. "d256" indicates capacity-enhanced runs.

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

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	Med. 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

Table 10. Emergence timeline for $\mathbb{F}_2[\mathbb{Z}_6]$, whose Wedderburn decomposition has semisimple quotient $\mathbb{F}_2 \times \mathbb{F}_4$ and radical fraction $1/2$ (seed 42). The $\mathbb{F}_2[\mathbb{Z}_2]$ factor contains the radical; the $\mathbb{F}_2[\mathbb{Z}_3] \cong \mathbb{F}_2 \times \mathbb{F}_4$ factor is semisimple.

Component	IIA > 0.8	Stable	Final
$\mathbb{Z}_2$ (radical, 4 elem.)	step 1K	19K	0.979
$\mathbb{Z}_3$ (semisimple, 8 elem.)	step 6K	12K	1.000
Grokking (test_acc ≥ 0.95)	step 19.2K		

Radical component algorithms. Our framework verifies that grokked models *decompose* computation along Wedderburn components, but does not characterize the algorithm within each component. For semisimple components, two-layer models converge to Fourier-based algorithms (logit cosine similarity ≥ 0.976; Appendix D). For radical components, the internal algorithm remains unknown: logit similarity is low ($\sim$0.12 for one-layer models), and the (WC3) condition cannot be meaningfully verified. Identifying what computation radical components perform—and whether it has a clean algebraic interpretation—is an important open question.

6.1. Limitations

1. **Mostly commutative algebras.** Complete Wedderburn alignment (all components) is verified only for commutative algebras. Non-commutative coverage is limited to $\mathbb{F}_2[S_3]$ (modular) and S_5 group multiplication (semisimple, selective alignment; Section 4.5).
2. **Scale.** The largest algebra that successfully grokked has 512 elements (343 for cross-characteristic). Algebras with ≥ 1024 elements (Table 8) fail to grok at any tested capacity, likely due to model capacity limitations.
3. **Radical coverage.** We test $n = 6$ radical-containing commutative algebras with radical fractions up to 1/2, plus $\mathbb{F}_2[S_3]$ (non-commutative, modular). Algebras with higher radical fractions (e.g., $\mathbb{F}_2[\mathbb{Z}_4 \times \mathbb{Z}_2]$, rf $= 7/8$) grok but have only a single Wedderburn block, precluding IIA decomposition.
4. **Architecture specificity.** All experiments use 1–2 layer Transformers. Whether deeper or different architectures change the results is unknown.
5. **Patching layer.** All IIA measurements patch at the embedding layer (`hook_resid_pre_0`). Whether Wedderburn alignment is maintained through subsequent attention and MLP layers has not been verified.

7. 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 but apply IIA over $\mathbb{F}_7$ rather than Fourier analysis, enabling a causal (rather than correlational) test of component independence. Zhong et al. (2023) identified "clock" and "pizza" algorithms for modular arithmetic. Stander et al. (2024) analyzed coset-based algorithms for group multiplication. Our work unifies these findings under the Wedderburn framework and provides the first causal evidence (via IIA) for component independence.

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 characterizes non-commutativity as a capacity threshold rather than an absolute barrier.

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.

8. 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. IIA verification across 11 commutative algebras (5 semisimple, 6 radical-containing; char 2/3/5/7), 1 non-commutative algebra ($\mathbb{F}_2[S_3]$), and 1 non-commutative group (S_5) shows that grokked models consistently decompose computation along Wedderburn components, with error compartmentalization ($3.8\times$ chance) confirming that component independence is structural. These results establish the Wedderburn decomposition as a structural invariant of grokking and open questions about the optimization dynamics that drive alignment.

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.