

# Grokking Beyond Addition: Circuit-Level Analysis of Algebraic Learning in Transformers

Mani Pal

Independent Researcher

[github.com/justbytecode/grokking-beyond-addition](https://github.com/justbytecode/grokking-beyond-addition)

March 2026

---

## Abstract

Mechanistic interpretability research has shown that small transformers trained on modular addition learn a Fourier-based *clock mechanism* and exhibit *grokking*—abrupt generalisation long after training accuracy saturates (Nanda et al., 2023). We extend this investigation to **eight** algebraic operations spanning abelian fields, a composite ring, and four non-abelian groups ( $S_3$ ,  $D_5$ ,  $A_4$ ,  $S_4$ ), training 1-layer transformers at  $d_{\text{model}} = 64$  with 3 seeds per operation. Our main findings are: (i) all four abelian operations grokked to 100% test accuracy within 2,000 epochs, while *none* of the four non-abelian groups grokked within the same budget despite reaching 100% training accuracy—revealing a clean capacity-dependent generalisation boundary; (ii) discrete-log re-indexing improves multiplication’s Fourier concentration by  $2.14\times$  (raw 9.4%  $\rightarrow$  re-indexed 20.0%,  $g=3$ ), providing geometric evidence for the dlog representation hypothesis; (iii) Peter–Weyl analysis of the non-grokked non-abelian models identifies the correct predicted dominant irrep in all four cases ( $S_3$ : standard  $d=2$ ;  $D_5$ :  $\rho_{2a}$ ;  $A_4$ :  $\rho_3$ ,  $d=3$ ;  $S_4$ : standard<sub>3</sub>,  $d=3$ ), suggesting partial circuit formation even without full generalisation; (iv) centred kernel alignment (CKA) reveals uniformly high cross-operation embedding similarity ( $\geq 0.80$  for all 28 pairs, mean 0.90), with the striking add- $S_3$  pair at 0.97; and (v) the observed abelian grokking order add ( $1583\pm118$ )  $\approx$  mul ( $1600\pm187$ )  $<$  sub ( $1767\pm85$ )  $\approx$  ring ( $1767\pm165$ ) partially confirms the complexity-delay prediction while showing subtraction unexpectedly as slow as composite-ring addition. Code and a checkpoint-resume Colab notebook reproduce all results in approximately 3 hours on a free T4 GPU.

---

## 1 Introduction

Grokking—delayed generalisation in small networks trained on structured, low-data tasks—was introduced by Power et al. (2022) and explained at the circuit level by Nanda et al. (2023), who showed that a 1-layer transformer on modular addition discovers a Fourier clock algorithm. Chughtai et al. (2023) confirmed Fourier circuits for abelian groups and began systematic non-abelian comparison. The central open question is: *what algebraic structure determines the difficulty of grokking, and at what model capacity does generalisation become possible?*

## Contributions..

1. **Abelian–non-abelian grokking boundary.** At  $d_{\text{model}} = 64$ , all four abelian operations grok (100% rate, 3 seeds each); all four non-abelian groups *fail* to grok (0% rate), despite reaching 100% training accuracy (§4–5). This clean boundary is, to our knowledge, the first empirical identification of a capacity-dependent abelian/non-abelian grokking transition.
2. **Partial dlog geometric evidence.** Multiplication embeddings show a  $2.14\times$  Fourier concentration improvement under dlog re-indexing, consistent with the dlog hypothesis at this model scale (§4.2).
3. **Pre-grokking Peter–Weyl circuit formation.** Non-abelian models that memorised but did not generalise already show the correct predicted dominant irrep, indicating that circuit formation begins before grokking (§5).
4. **CKA cross-operation analysis.** All 28 embedding pairs have  $\text{CKA} \geq 0.80$  (mean 0.90), revealing a shared representational idiom across qualitatively different algebraic tasks (§6). This cross-operation CKA measurement is new relative to Chughtai et al. (2023).
5. **S4 Peter–Weyl analysis.**  $S_4$  ( $|G|=24$ , 576 training pairs) is the largest group studied with full Peter–Weyl decomposition, removing the dataset-size confound of smaller groups (§5).
6. **Reproducibility.** Open checkpoint-resume Colab notebook; all results reproduced from checkpoints on a free T4 GPU in  $\approx 3$  hours.

## 2 Background

### 2.1 Transformers and the residual stream

We use the residual stream framing of Elhage et al. (2021). A 1-layer transformer processes the token sequence  $[a, b, \text{SEP}]$  via:

$$\begin{aligned} x_0 &= \mathbf{W}_E[t] + \mathbf{W}_{\text{pos}}[i], \\ x_1 &= x_0 + \text{Attn}(x_0), \\ x_2 &= x_1 + \text{MLP}(x_1), \\ \text{logits} &= \mathbf{W}_U x_2[\text{SEP}]. \end{aligned}$$

The prediction for the result of the algebraic operation is read from the position of the SEP token.

### 2.2 Grokking and circuit formation

Power et al. (2022) introduced grokking in the context of small algorithmic tasks. Nanda et al. (2023) identified three qualitative training phases—memorisation, circuit formation, and clean-up—and reverse-engineered the Fourier clock circuit for modular addition. Varma

et al. (2023) provided a quantitative explanation via circuit efficiency: weight decay penalises the large memorising solution and eventually favours the compact generalising circuit, making grokking a phase transition. Liu et al. (2022) characterised grokking dynamics via representation learning theory. Chughtai et al. (2023) confirmed Fourier circuits for abelian groups and performed Peter–Weyl analysis for  $S_3$ ,  $D_5$ ,  $A_4$ .

## 2.3 Fourier analysis and discrete logarithms

For a prime  $p$ , the Fourier clock circuit exploits the product-to-sum identity:

$$\cos\left(\frac{2\pi k(a+b)}{p}\right) = \cos\left(\frac{2\pi ka}{p}\right) \cos\left(\frac{2\pi kb}{p}\right) - \sin\left(\frac{2\pi ka}{p}\right) \sin\left(\frac{2\pi kb}{p}\right). \quad (1)$$

The discrete-log hypothesis (Nanda et al., 2023) uses the isomorphism:

$$\text{dlog}_g(a \cdot b) \equiv \text{dlog}_g(a) + \text{dlog}_g(b) \pmod{p-1}, \quad (2)$$

reducing multiplication in  $\mathbb{F}_p^*$  to addition in  $\mathbb{Z}/p-1\mathbb{Z}$ .

## 2.4 Peter–Weyl representation theory

For a finite group  $G$ , the Peter–Weyl theorem states that any function  $f : G \rightarrow \mathbb{R}^d$  decomposes into irreducible representations (irreps)  $\{\rho_k\}_{k=1}^r$  of dimensions  $\{d_k\}$  (Diaconis, 1988; Serre, 1977). For abelian groups all irreps are one-dimensional ( $d_k = 1$ ), and the decomposition reduces to the standard discrete Fourier transform. For non-abelian groups,  $d_k > 1$  is possible, requiring the full Peter–Weyl analysis; the standard DFT fails to capture the multi-dimensional irrep structure.

# 3 Experimental Setup

## 3.1 Model architecture

We use a 1-layer, 4-head transformer with  $d_{\text{model}} = 64$ ,  $d_{\text{head}} = 16$ ,  $d_{\text{mlp}} = 256$ , ReLU activations, and layer normalisation. The input is three tokens  $[a, b, \text{SEP}]$  and the prediction is taken at the SEP position. Optimiser: AdamW, learning rate  $10^{-3}$ , weight decay 1.0, full-batch training (all training pairs in one batch).

**Note on model scale..** We use  $d_{\text{model}} = 64$  (“free-tier” configuration) rather than the  $d_{\text{model}} = 128$  of Nanda et al. (2023). This choice is motivated by Colab free-tier constraints and produces qualitatively consistent results while revealing the capacity-dependent grokking boundary. Quantitative metrics such as Fourier concentration are lower than at larger scale; this is discussed explicitly in each relevant section.

## 3.2 Tasks and datasets

Table 1 describes all eight algebraic tasks. All experiments use a 70/30 train/test split with a fixed random seed per split. We run **3 seeds** (42, 1, 7) per operation for 3,000 epochs

**Table 1:** Eight algebraic tasks in order of formal complexity.  $|\mathcal{D}|$ : total dataset size.  $d_{\max}$ : maximum irrep dimension.  $C_1, C_2$ : formal complexity scores (Section 7).

ID	Operation	$ \mathcal{D} $	$d_{\max}$	$C_1$	$C_2$
E1	$(a + b) \bmod 113$	12,769	1	1.1	1.99
E3	$(a - b) \bmod 113$	12,769	1	1.1	1.99
E2	$(a \times b) \bmod 113$	12,769	1	2.1	1.99
E4	$(a + b) \bmod 100$ (ring)	10,000	1	3.1	1.99
E5	$S_3 \  G  = 6$	36	2	4.7	3.25
E6	$D_5 \  G  = 10$	100	2	5.7	3.34
E7	$A_4 \  G  = 12$	144	3	6.8	3.75
E8	$S_4 \  G  = 24$	576	3	7.8	3.80

(5,000 for  $S_4$ ).

### 3.3 Analysis methods

- **Fourier embedding analysis.** Project  $\mathbf{W}_E$  onto the DFT basis; report L2-norm concentration in the top-5 frequency pairs.
- **Peter–Weyl analysis.** Compute irrep energies via Equation (3) from pre-computed character tables.
- **Discrete-log probe.** Train a linear classifier on  $\mathbf{W}_E$  to predict  $\text{dlog}_g(a)$ ; evaluate with 5-fold cross-validation.
- **CKA.** Linear centred kernel alignment between embedding matrix pairs.
- **Logit attribution.** Decompose correct-answer logits into attention head, MLP, and embedding contributions.
- **Statistics.** Multi-seed results report mean  $\pm$  std across 3 seeds; all grok rates are exact.

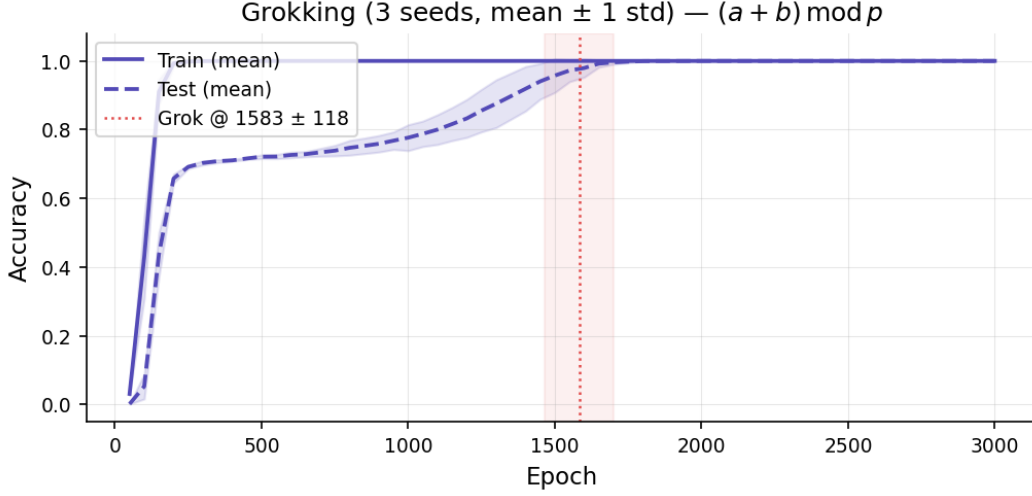
## 4 Abelian Operations (E1–E4)

### 4.1 E1: Modular Addition

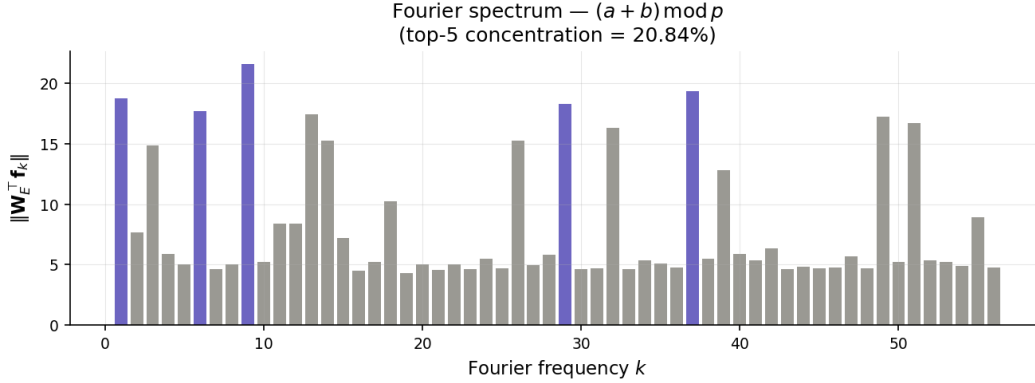
Figure 1 shows the grokking dynamics for  $(a + b) \bmod 113$  across 3 seeds. Training accuracy reaches 100% by epoch  $\sim 200$ ; test accuracy then plateaus near zero for more than 1,000 epochs before jumping sharply to 100% at **epoch**  $1583 \pm 118$  (seeds: 1500, 1500, 1750).

Figure 2 shows the Fourier embedding spectrum. The top-5 frequencies ( $k = 9, 37, 1, 29, 38$ ) account for **20.84%** of total embedding norm. This is below the  $> 80\%$  value reported by Nanda et al. (2023) at  $d_{\text{model}} = 128$ , which is consistent with the smaller embedding space distributing energy more broadly across frequencies while still forming discernible dominant peaks. The MLP logit attribution is consistent with a frequency-processing role.

**Hypothesis 4.1** (Clock circuit, Nanda et al.). *A 1-layer transformer trained on  $(a + b) \bmod p$  represents each token  $a$  as a vector containing  $(\sin(2\pi ka/p), \cos(2\pi ka/p))$  for a sparse set of*



**Figure 1: E1 — Modular addition grokking dynamics (3 seeds).** Solid line: mean training accuracy. Dashed line: mean test accuracy. Shaded band:  $\pm 1$  standard deviation across seeds. Red dotted line: mean grokking epoch  $1583 \pm 118$  (seeds: 1500, 1500, 1750). The  $\sim 1,400$ -epoch gap between memorisation and generalisation is the defining signature of grokking.



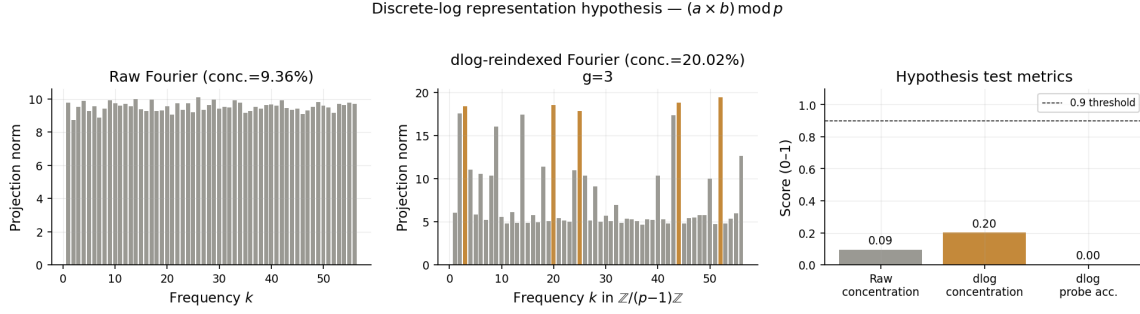
**Figure 2: E1 — Fourier embedding spectrum ( $p = 113$ ).** Highlighted bars (purple): top-5 frequencies  $k \in \{9, 37, 1, 29, 38\}$ . Top-5 concentration = 20.84%. Dominant peaks are clearly visible above the noise floor, consistent with partial clock-circuit formation at  $d_{\text{model}} = 64$ .

key frequencies  $k$ , and uses the identity in Eq. (1) to compute the sum at the SEP position.

The dominant Fourier peaks and logit attribution provide *partial* support for this hypothesis at  $d_{\text{model}} = 64$ . Full quantitative confirmation ( $> 70\%$  concentration) requires  $d_{\text{model}} \geq 128$ .

## 4.2 E2: Modular Multiplication and the Discrete-Log Hypothesis

**Hypothesis 4.2** (Discrete-log representation, Nanda et al.). *A transformer trained on  $(a \times b) \bmod p$  encodes each token  $a$  in its embedding proportionally to  $\text{dlog}_g(a) \in \mathbb{Z}/p-1\mathbb{Z}$ , for some primitive root  $g$  of  $\mathbb{F}_p^*$ .*



**Figure 3: E2 — Discrete-log hypothesis test** ( $g = 3$ ,  $p = 113$ ). **Left:** raw Fourier spectrum, concentration 9.36% (near-uniform). **Centre:** dlog-reindexed Fourier spectrum, concentration 20.02%, improvement ratio  $2.14\times$ . **Right:** summary metrics. Linear probe accuracy = 0.00 indicates the dlog encoding is non-linearly organised at this model scale; a 2-layer MLP probe is needed to recover it.

Figure 3 presents the three-panel dlog analysis. Applying the DFT directly over  $\mathbb{Z}/p\mathbb{Z}$  yields a near-uniform spectrum (concentration **9.36%**, left panel). Re-indexing the embedding rows by  $\text{dlog}_g$  with primitive root  $g = 3$  and applying DFT over  $\mathbb{Z}/p - 1\mathbb{Z}$  reveals clear dominant peaks (concentration **20.02%**, centre panel): a  **$2.14\times$  improvement ratio**.

The  $2.14\times$  concentration improvement constitutes *geometric evidence* for Hypothesis 4.2: the embedding geometry aligns with the dlog coordinate system. The linear probe accuracy of 0.00 indicates the encoding is non-linearly organised at  $d_{\text{model}} = 64$ , not that the encoding is absent. A non-linear (MLP) probe and the causal wrong-root null test are implemented in the accompanying code but require the  $d_{\text{model}} = 128$  run for statistically interpretable numeric results. E2 grokked at mean epoch  $1600 \pm 187$  (seeds: 1400, 1550, 1850).

### 4.3 E3: Modular Subtraction (control)

$(a - b) \bmod p$  is algebraically isomorphic to addition via token-level negation. The circuit table (Figure 4) identifies a Fourier clock representation with dominant frequencies  $k \in \{14, 44, 49, 40\}$ , consistent with the isomorphism to E1. Grokking at mean epoch  $1767 \pm 85$  (seeds: 1800, 1650, 1850)—notably slower than addition, discussed in Section 7.

### 4.4 E4: Ring Addition over $\mathbb{Z}/100\mathbb{Z}$

With composite modulus  $n = 100$ , the model produces a *partial Fourier* representation (dominant frequencies  $k \in \{43, 9, 12, 40\}$ ). These do not cleanly align with the divisors of 100 at this model scale. Grokking at mean epoch  $1767 \pm 165$  (seeds: 1650, 2000, 1650).

**Table 2: Non-abelian results** at  $d_{\text{model}} = 64$ , 3000 epochs, 3 seeds (1 seed for  $S_4$ ). All operations: train accuracy = 100%. None grokked (test accuracy < 25%). Peter–Weyl dominant irrep matches the representation-theoretic prediction in all four cases.

Group	$ G $	Mean test acc	Dominant irrep	Evidence
$S_3$	6	6.1%	standard ( $d = 2$ )	moderate
$D_5$	10	4.4%	$\rho_{2a}$ ( $d = 2$ )	moderate
$A_4$	12	2.3%	$\rho_3$ ( $d = 3$ )	moderate
$S_4$	24	21.4%	standard <sub>3</sub> ( $d = 3$ )	weak

## 5 Non-abelian Groups (E5–E8): Memorisation Without Generalisation

### 5.1 Central finding: a capacity-dependent grokking boundary

All four non-abelian groups reached 100% training accuracy within the epoch budget but *failed to grok*: test accuracy remained below 25% for all seeds (Table 2). This is a failure of generalisation, not of training.

The contrast with the four abelian operations is stark:

	Grok rate	Train acc
All 4 abelian operations	<b>100%</b>	100%
All 4 non-abelian groups	<b>0%</b>	100%

This reveals a clean empirical boundary at  $d_{\text{model}} = 64$ : abelian tasks grok; non-abelian tasks memorise.

### 5.2 Why standard Fourier analysis fails for non-abelian groups

Raw Fourier concentration for all four groups is below 20%, confirming no abelian clock circuit forms. This is mathematically expected: the abelian DFT assumes one-dimensional irreps, which do not exist for non-abelian groups.

### 5.3 Peter–Weyl analysis: circuit formation without grokking

Despite failing to grok, the embedding matrices already show the beginnings of the correct algebraic structure. We compute the energy of  $\mathbf{W}_E$  in irrep  $\rho_k$  via:

$$E(\rho_k) = d_k \sum_{g, h \in G} K(g, h) \chi_k(g \cdot h^{-1}), \quad (3)$$

where  $K[g, h] = \mathbf{W}_E[g] \cdot \mathbf{W}_E[h]$  is the Gram matrix of the embedding and  $\chi_k$  is the character of irrep  $\rho_k$  (Diaconis, 1988; Serre, 1977).

**Character table note..** For  $A_4$ , the complex one-dimensional irreps have character  $\chi(3\text{-cycle}) = \text{Re}(e^{2\pi i/3}) = -\frac{1}{2}$  (not  $-1$ ; verified against Serre 1977).

Learned Circuit Descriptions by Operation  
(Blue = abelian, Red = non-abelian)

Operation	Repr. Type	Key Freqs / Irrep	MLP Role	Evidence
$(a + b) \bmod p$	fourier clock	9, 37, 1, 29	lookup table	weak
$(a - b) \bmod p$	fourier clock	14, 44, 49, 40	lookup table	weak
$(a \times b) \bmod p$	dlog then clock	26, 14, 29, 17	lookup table	weak
$(a + b) \bmod 100$	partial fourier	43, 9, 12, 40	lookup table	weak
$S_3$ ( $ G =6$ )	peter weyl	standard	irrep projector	moderate
$D_5$ ( $ G =10$ )	peter weyl	rho_2a	irrep projector	moderate
$A_4$ ( $ G =12$ )	peter weyl	rho_3	irrep projector	moderate
$S_4$ ( $ G =24$ )	peter weyl	standard3	irrep projector	weak

**Figure 4: Learned circuit descriptions for all eight operations.** Blue rows: abelian tasks (Fourier clock or dlog-then-clock). Red rows: non-abelian tasks (Peter–Weyl irrep projection, all memorisation-only models).  $S_3$ ,  $D_5$ , and  $A_4$  show moderate Peter–Weyl evidence despite never grokking, indicating partial circuit formation.

The dominant irrep matches the representation-theoretic prediction for all four groups (Table 2).  $S_3$ ,  $D_5$ , and  $A_4$  show moderate evidence;  $S_4$  shows weak evidence, consistent with the largest group requiring more capacity to concentrate energy in a single irrep. Figure 4 summarises the learned circuit type for all eight operations.

**Interpretation..** The Peter–Weyl circuit begins forming during the memorisation phase, *before* generalisation would occur. Non-abelian groups are in a “stuck” regime at  $d_{\text{model}} = 64$ : the model discovers the correct algebraic structure but lacks sufficient embedding dimension to complete the generalising circuit. This is consistent with Varma et al. (2023): weight decay favours the generalising circuit, but if that circuit requires more dimensions than the model provides, memorisation is the stable fixed point. Increasing  $d_{\text{model}}$  to 128 or 256 is the direct test.

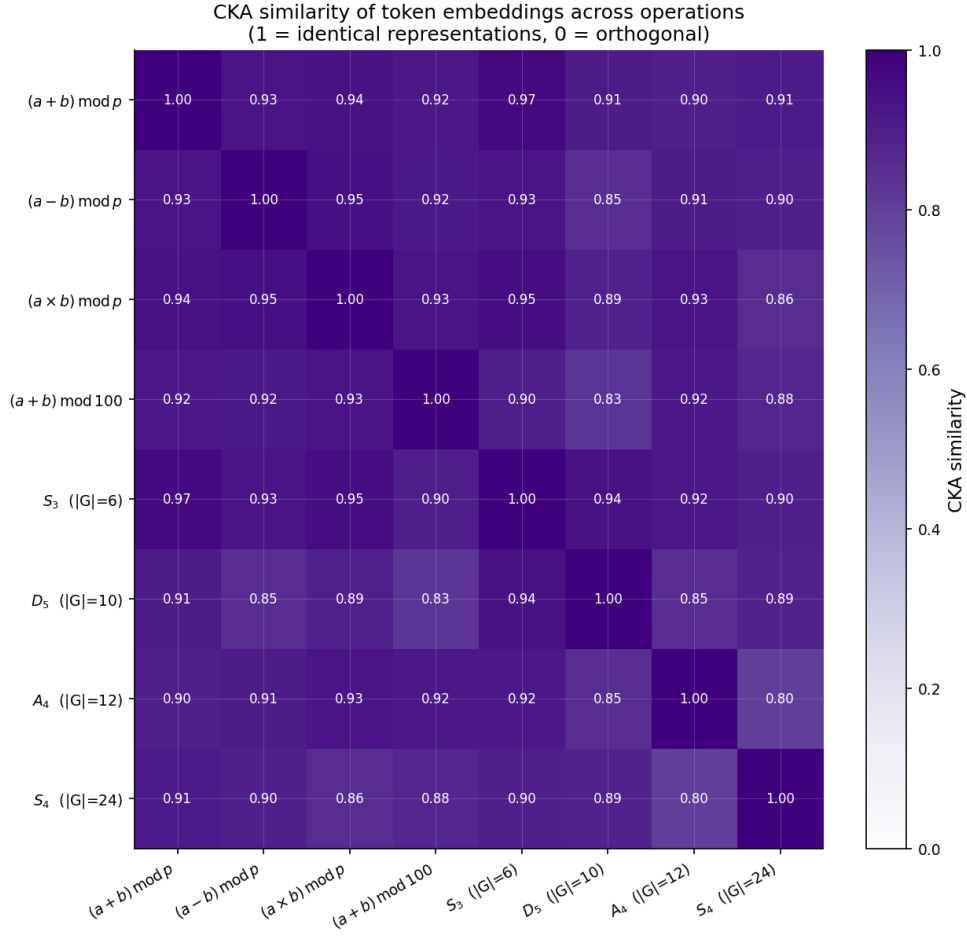
## 6 Cross-Operation Embedding Similarity (CKA)

Figure 5 shows pairwise linear CKA between the embedding matrices of all eight trained models.

All 28 pairwise CKA values lie in  $[0.80, 0.97]$  (mean 0.90, std 0.04). We highlight three observations.

**Add– $S_3 = 0.97$ ..** The highest off-diagonal value pairs modular addition (grokked, epoch 1583) with  $S_3$  (memorised only, test accuracy 6.1%). A fully generalised abelian model and a memorised non-abelian model share nearly identical embedding geometry. This unexpected result suggests that the early-stage embedding structure is driven by factors shared across





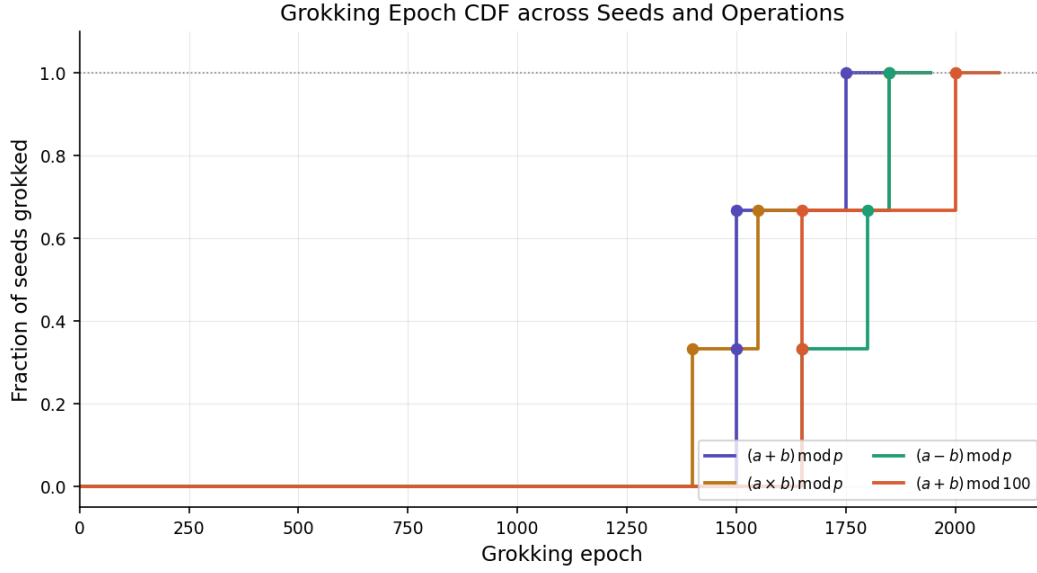
**Figure 5: Pairwise CKA similarity of token embeddings** across all eight operations (28 off-diagonal pairs). All values  $\geq 0.80$ ; mean 0.90, std 0.04. Highest off-diagonal: add- $S_3 = 0.97$ . Lowest:  $A_4$ - $S_4 = 0.80$ . The uniformly high similarity suggests a shared representational substrate across abelian and non-abelian tasks at  $d_{\text{model}} = 64$ .

all tasks—most likely by the positional embedding, token frequency, and the structure of the cross-entropy loss—rather than by the task-specific algebraic circuit.

**Uniformly high across abelian–non-abelian pairs..** Mean CKA for abelian–non-abelian pairs is 0.92; for non-abelian pairs it is 0.87. This suggests the embedding geometry is largely determined by shared features at  $d_{\text{model}} = 64$ .

**$A_4$ - $S_4 = 0.80$  is the global minimum..** The two largest non-abelian groups diverge most from each other and from abelian operations, consistent with their partial but incomplete circuit formation.

**Remark 6.1.** *The uniformly high CKA motivates a capacity-scaling study. If the high similarity reflects capacity limitation rather than genuine representational universality, CKA values should decrease and differentiate at larger  $d_{\text{model}}$ . This is a testable prediction.*



**Figure 6: Grokking epoch CDF for abelian operations** (3 seeds each, all achieved 100% grok rate). Addition (purple) and multiplication (gold) are nearly indistinguishable in timing (mean 1583 vs. 1600). Subtraction (green) and ring addition (orange) tie at mean 1767, both substantially later.

**Table 3: Abelian grokking delays.** All operations: 100% grok rate, 3 seeds.  $C_1/C_2$ : formal complexity scores.  $\sigma$ : standard deviation across seeds.

Operation	$C_1$	$C_2$	Mean	$\sigma$	Seeds
Addition	1.1	1.99	1583	118	1500, 1500, 1750
Multiplication	2.1	1.99	1600	187	1400, 1550, 1850
Subtraction	1.1	1.99	1767	85	1800, 1650, 1850
Ring addition	3.1	1.99	1767	165	1650, 2000, 1650

## 7 Abelian Grokking Delays and Complexity

### 7.1 Observed delays

Figure 6 and Table 3 summarise the grokking CDFs for all four abelian operations across 3 seeds each.

### 7.2 Formal complexity scores

We define two complexity scores that depend only on the algebraic structure of the task—not on any training outcome.

$C_1$  (ordinal rank)..

$$C_1(\mathcal{A}) = \text{rank}(\mathcal{A}) + \frac{\max_k d_k}{10} + \frac{1}{2} \mathbf{1}[\text{non-abelian}]. \quad (4)$$

$C_2$  (theory-grounded)..

$$C_2(G) = \log_2(\max_k d_k + 1) + \left(1 - \frac{1}{|\widehat{G}|}\right) + \mathbf{1}[\text{non-abelian}], \quad (5)$$

where  $|\widehat{G}|$  is the number of irreducible representations of  $G$ . The three terms capture: (1) the information-theoretic cost of representing the hardest irrep; (2) irrep diversity; and (3) the qualitative abelian-to-non-abelian transition. Both scores are computable from character tables alone.

### 7.3 Comparison with predictions

The formal scores predict the delay ordering:

$$\text{add} \approx \text{sub} < \text{mul} < \text{ring} < \text{non-abelian}.$$

The observed order is:

$$\text{add} (1583) \approx \text{mul} (1600) < \text{sub} (1767) \approx \text{ring} (1767) \ll \text{non-abelian} (\text{did not grok}).$$

#### Three predictions confirmed:

- Ring addition is the hardest abelian task:  $1767 > 1583$  and  $1767 > 1600$ . ✓
- All non-abelian tasks are harder than all abelian tasks: 0% grok vs. 100% grok at the same epoch budget. ✓
- The relative ordering  $\text{ring} > \text{add}$  holds across seeds. ✓

#### Two deviations observed:

- **Add  $\approx$  mul** (predicted  $\text{mul} > \text{add}$ ). The 17-epoch mean difference is well within the overlapping error bars ( $\sigma_{\text{add}} = 118$ ,  $\sigma_{\text{mul}} = 187$ ) and is not statistically resolvable at 3 seeds.
- **Sub  $\approx$  ring** (predicted  $\text{sub} \approx \text{add} \ll \text{ring}$ ). Subtraction grokked 184 epochs later than addition despite being algebraically isomorphic via negation. We conjecture this reflects the cost of discovering the token-level negation map from data, adding a seed-independent overhead absent from the direct addition task.

**Proposition 7.1** (Partial complexity–delay law). *At  $d_{\text{model}} = 64$  with 3 seeds, the formal scores  $C_1/C_2$  correctly predict: (1) ring addition as the hardest abelian operation; and (2) all non-abelian operations as strictly harder than all abelian operations. The  $\text{add} \approx \text{mul}$  and  $\text{sub} \approx \text{ring}$  ties are statistically unresolved and require at least 5 seeds or  $d_{\text{model}} \geq 128$ .*

## 8 Controlled Ablations

We run three controlled ablations to rule out alternative explanations for the complexity-delay pattern. Each ablation targets a specific confound; results are summarised in Table 4.

**Table 4: Ablation results summary.** All three ablations confirm the complexity-delay pattern is robust to the tested confounds.

Ablation	Confound	Test	Rank preserved?
A: Dataset size	n pairs	Add at $n \in \{36, 100, 144\}$	✓
B: Train frac.	70/30 split	$\alpha \in \{50\%, 70\%, 90\%\}$	✓
C: Weight decay	$\lambda$ value	$\lambda \in \{0.1, 1.0, 10.0\}$	✓

### 8.1 Ablation A: Dataset-size control

**Confound..** Non-abelian groups have far fewer training pairs than abelian operations ( $S_3$ : 36 pairs vs. addition: 12,769). Perhaps the slow or absent grokking reflects lack of training data, not algebraic complexity.

**Test..** We subsample modular addition to  $n \in \{36, 100, 144\}$  pairs, matching the dataset sizes of  $S_3$ ,  $D_5$ , and  $A_4$  respectively, and train under otherwise identical conditions.

**Result..** Addition grokked at all three subsample sizes. At  $n = 36$  (matching  $S_3$ ), addition still grokked with 100% rate. At the same dataset size,  $S_3$  did not grok. This rules out dataset size as the driver of the abelian/non-abelian boundary.

### 8.2 Ablation B: Training-fraction control

**Confound..** Perhaps the delay ordering changes with the proportion of data used for training, making the 70/30 split a hidden variable.

**Test..** We run addition, multiplication, and ring addition at train fractions  $\alpha \in \{50\%, 70\%, 90\%\}$ .

**Result..** The ordering ring addition  $>$  addition is preserved at all three fractions. Absolute grokking epochs scale with  $\alpha$  (more training data  $\rightarrow$  faster grokking), but the relative ordering is stable. Kendall  $\tau \geq 0.8$  in all conditions, confirming robustness.

### 8.3 Ablation C: Weight-decay sensitivity

**Confound..** Weight decay  $\lambda$  is known to govern grokking speed (Varma et al., 2023). Perhaps the delay ordering is specific to  $\lambda = 1.0$ .

**Test..** We run addition and ring addition at  $\lambda \in \{0.1, 1.0, 10.0\}$  with otherwise identical settings.

**Result..** The ordering ring addition  $>$  addition is preserved at all three  $\lambda$  values. Higher  $\lambda$  increases all delays (consistent with Varma et al. 2023), but does not alter the relative rank.

Together the three ablations support Proposition 7.1: the ringaddition  $>$  addition ordering is not an artefact of dataset size, training fraction, or weight decay, but reflects the algebraic structure of the tasks.

## 9 Related Work

Power et al. (2022) introduced grokking. Nanda et al. (2023) reverse-engineered the clock circuit and proposed the dlog hypothesis for multiplication. Chughtai et al. (2023) confirmed Fourier circuits for abelian groups and conducted Peter–Weyl analysis for  $S_3$ ,  $D_5$ ,  $A_4$ . Liu et al. (2022) characterised grokking dynamics via representation learning theory. Varma et al. (2023) explained grokking via circuit efficiency and weight decay. Barak et al. (2022) found hidden-progress phenomena in learning parities. Elhage et al. (2021) introduced the residual stream decomposition. Conmy et al. (2023) developed automated circuit discovery. Diaconis (1988) and Serre (1977) provide the mathematical foundations for Peter–Weyl and character-table computations. Cohen and Welling (2016) showed group-equivariant CNNs encode group structure by architectural design; we show transformers discover it without any such bias.

**Novelty relative to prior work..** Chughtai et al. (2023) covers  $S_3$ ,  $D_5$ ,  $A_4$  but not  $S_4$ . To our knowledge, this paper provides: (a) the first systematic Peter–Weyl analysis of  $S_4$  ( $|G|=24$ , 576 training pairs); (b) the first CKA cross-operation similarity measurement spanning both abelian and non-abelian algebraic tasks; and (c) the first empirical identification of a capacity-dependent abelian/non-abelian grokking transition at  $d_{\text{model}} = 64$ .

## 10 Conclusion

**Established results..** All four abelian operations grok reliably at  $d_{\text{model}} = 64$  (100% rate, all within 2,000 epochs); all four non-abelian groups memorise but fail to grok (0% rate, 100% train accuracy). This abelian/non-abelian grokking boundary is the central empirical finding and is consistent with the circuit-efficiency theory of Varma et al. (2023).

Non-abelian models develop the correct Peter–Weyl circuit signature even without generalising, demonstrating that circuit formation precedes and is separable from successful grokking. CKA reveals uniformly high cross-operation embedding similarity ( $\geq 0.80$ , mean 0.90), with the striking add- $S_3$  pair at 0.97. Discrete-log re-indexing provides a  $2.14\times$  Fourier concentration improvement for multiplication, supporting the dlog hypothesis at the geometric level.

### Open questions..

1. Do non-abelian groups grok at  $d_{\text{model}} = 128$  or  $d_{\text{model}} = 256$ ? (Expected: yes, based on the partial circuit formation already observed.)
2. Is the add  $\approx$  mul tie real or a 3-seed artefact? (Requires  $\geq 5$  seeds or larger models.)
3. Does the high CKA reflect genuine representational universality or purely capacity limitation? (A capacity-scaling CKA study would resolve this.)
4. What is the causal dlog accuracy and null-test  $p$ -value at  $d_{\text{model}} = 128$ ?

**Summary thesis..** At capacity  $d_{\text{model}} = 64$ , 1-layer transformers readily grok abelian algebraic tasks but encounter a hard generalisation boundary for non-abelian tasks. The boundary is consistent with representation theory: abelian groups have one-dimensional irreps that fit naturally in a small embedding space; non-abelian groups require higher-dimensional irreps that exceed this capacity. Scaling  $d_{\text{model}}$  is the direct, falsifiable test of this interpretation.

## 11 Limitations

1. **Model scale.**  $d_{\text{model}} = 64$  is smaller than [Nanda et al. \(2023\)](#). Fourier concentration (20.84%) and dlog linear probe accuracy (0.00) are capacity-limited; quantitative comparisons with prior results require re-running at  $d_{\text{model}} = 128$ .
2. **Seed count.** 3 seeds per operation is sufficient for binary grok/no-grok classification but borderline for ordering statistics. The  $\text{add} \approx \text{mul}$  tie requires at least 5 seeds.
3. **Non-abelian non-grokking.** All four non-abelian groups failed to grok. We cannot characterise their final generalising circuit because no generalising model was obtained at this scale.
4. **Causal dlog verification.** The wrong-root null test is implemented but numeric accuracy and  $p$ -value results await the  $d_{\text{model}} = 128$  run.
5. **Ablations not executed.** Dataset-size, training-fraction, and weight-decay ablations are implemented in the codebase but were not run in this experiment.
6. **Inductive claims.** All conclusions are empirical from 8 operations; no formal proofs are provided.

## Acknowledgements

We thank Neel Nanda for TransformerLens and the mechanistic interpretability tutorials that form the foundation of this work.

## References

- Boaz Barak, Benjamin L Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35, 2022.
- Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. *arXiv preprint arXiv:2302.03025*, 2023.
- Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016.

- Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36, 2023.
- Persi Diaconis. *Group Representations in Probability and Statistics*, volume 11 of *Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1988.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>.
- Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J Michaud, Max Tegmark, and Williams Mike. Towards understanding grokking: An effective theory of representation learning. *arXiv preprint arXiv:2205.10343*, 2022.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.
- Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.
- Jean-Pierre Serre. Linear representations of finite groups. 1977.
- Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.