Understanding and acceleration of grokking phenomena in learning arithmetic operations via Kolmogorov-Arnold representation

2024 KIAS CAINS Fall Workshop

(Joint work with Yeoneung Kim, Minseok Kim)

Park, Yeachan

Korea Institute for Advanced Study

ychpark@kias.re.kr

November 21, 2024

Table of Contents

I. Introduction
- (i) Discovery of Grokking
- (ii) Possible explanation of Grokking

II. Understanding Grokking in modular arithmetic
- (i) Motivation

III. Kolmogorov-Arnold representation and modular arithmetic

IV. Acceleration of Grokking

V. Experiments

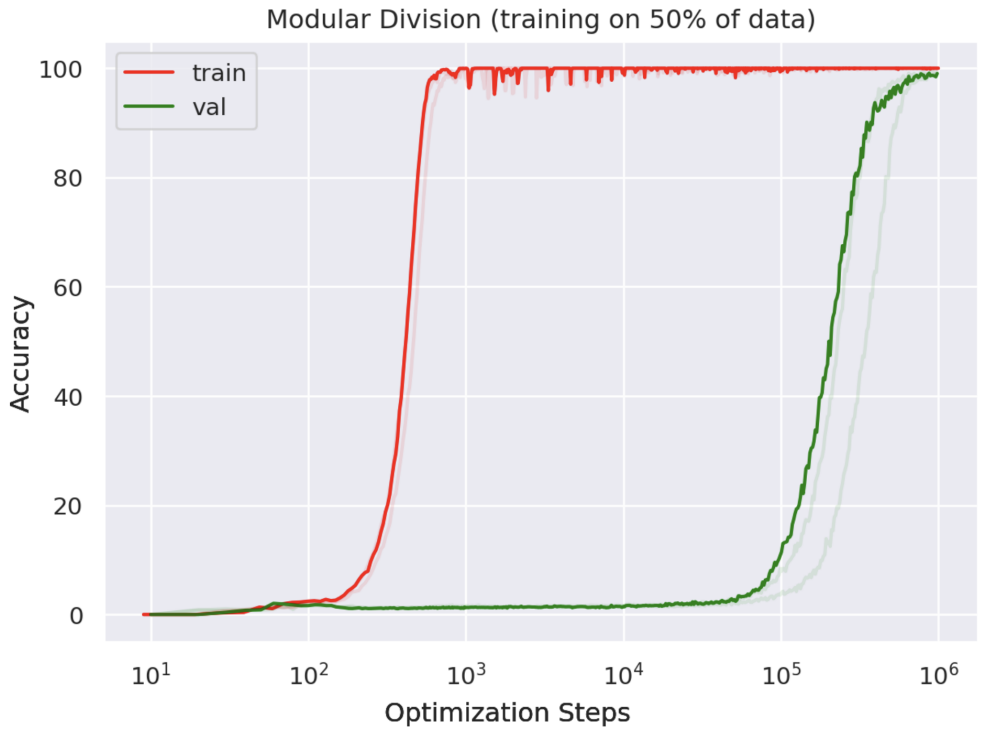
Introduction

Discovery of Grokking

Motivation

Objective:

$$12 \quad / \quad 6 \quad = \quad ? \ (\mathrm{mod}\ p) \tag{1}$$

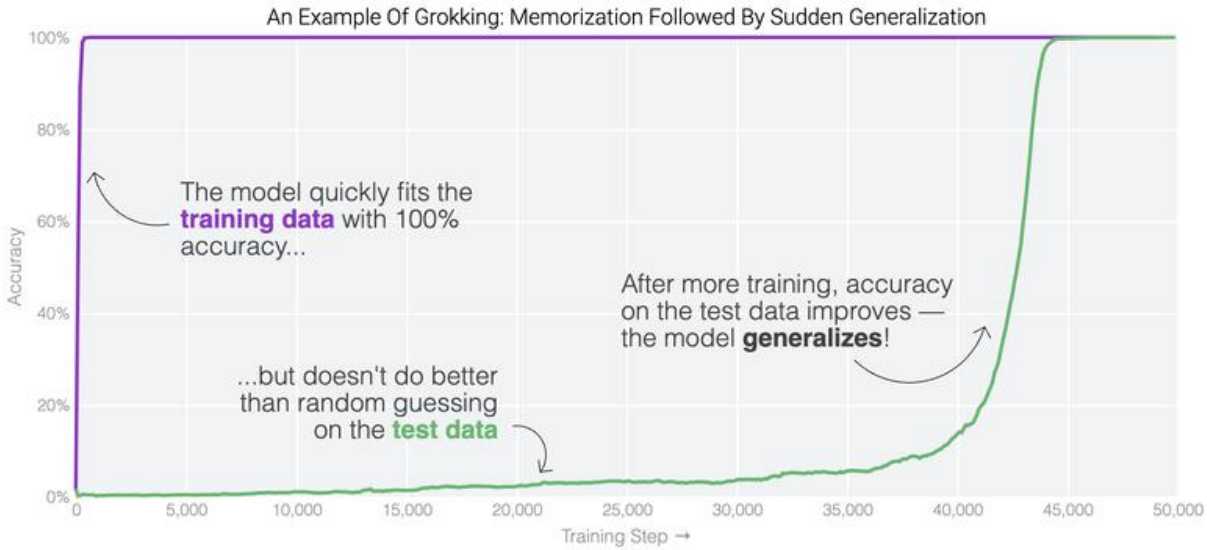


[1]

- Grokking phenomenon
- Sudden generalization after long training

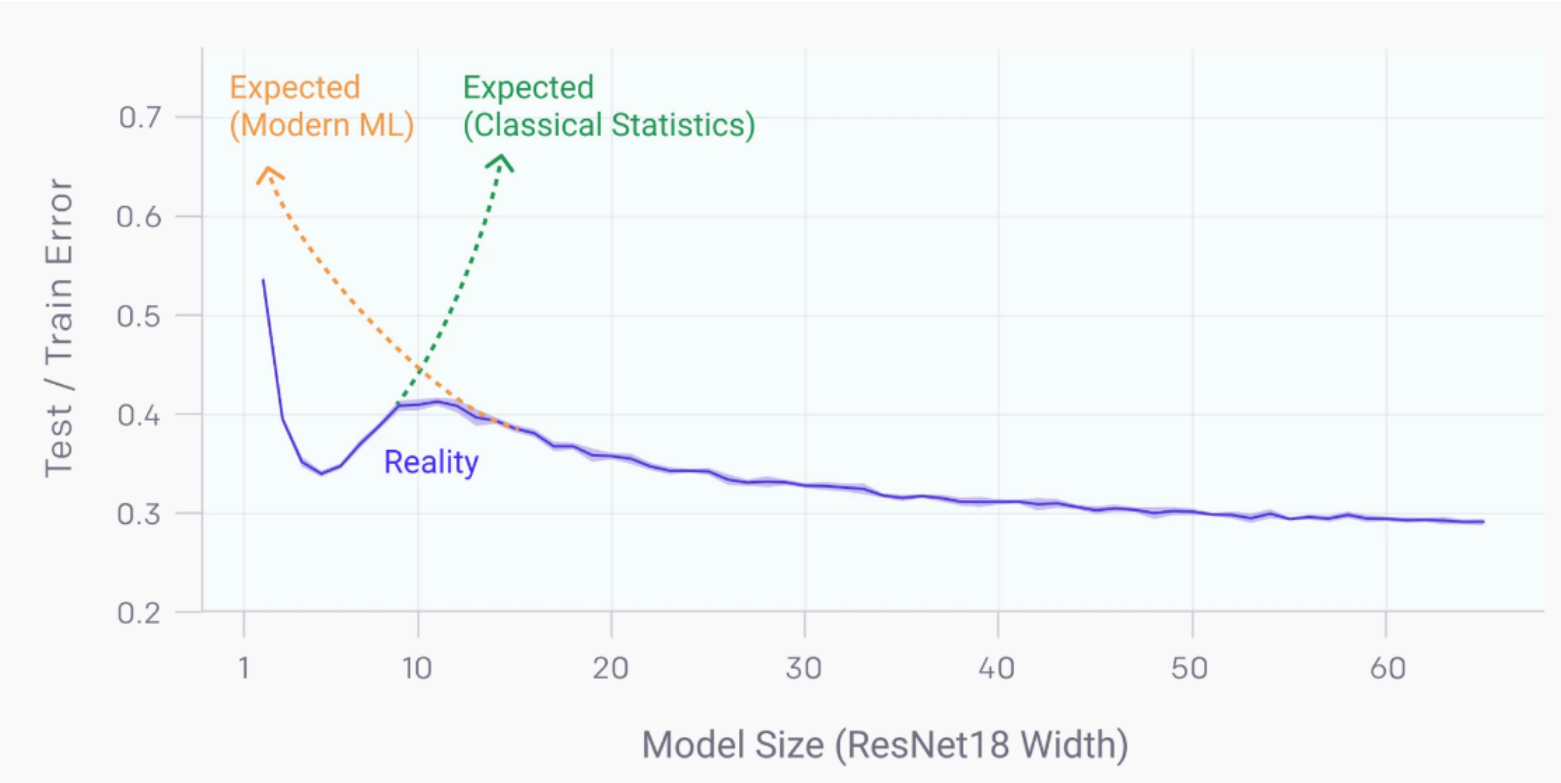
[1] Power, Alethea, et al. "Grokking: Generalization beyond overfitting on small algorithmic datasets." arXiv preprint arXiv:2201.02177 (2022).

Grokking?



- Memorize? Generalize?
- Should we train longer?

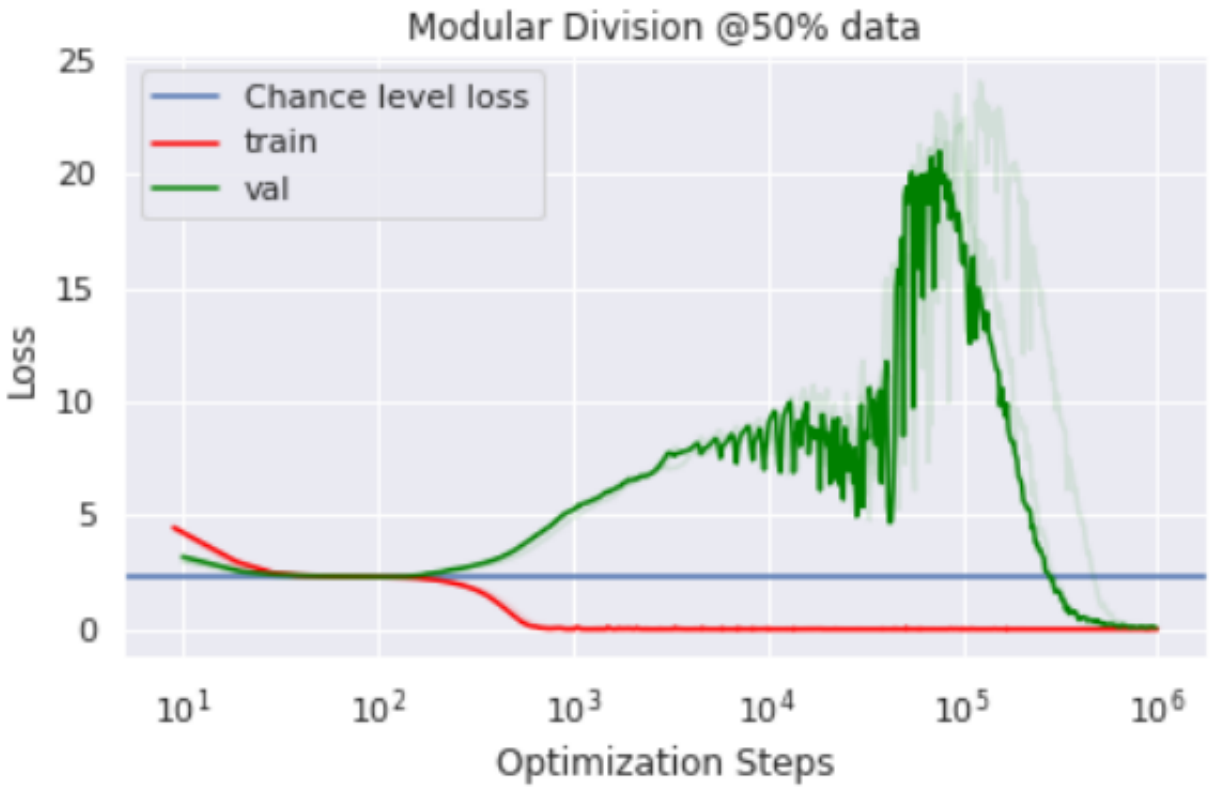
Deep double descent



[2]

[2] Nakkiran, Preetum, et al. "Deep double descent: Where bigger models and more data hurt." Journal of Statistical Mechanics: Theory and Experiment 2021.12 (2021): 124003.

Deep double descent in iteration



- Benign overfitting?
- Is overfitting desirable?

Possible explanation of Grokking

Explanation : large initialization and weight decay

- Large initialization induces grokking.
- Goldilocks zone: Optimal region.

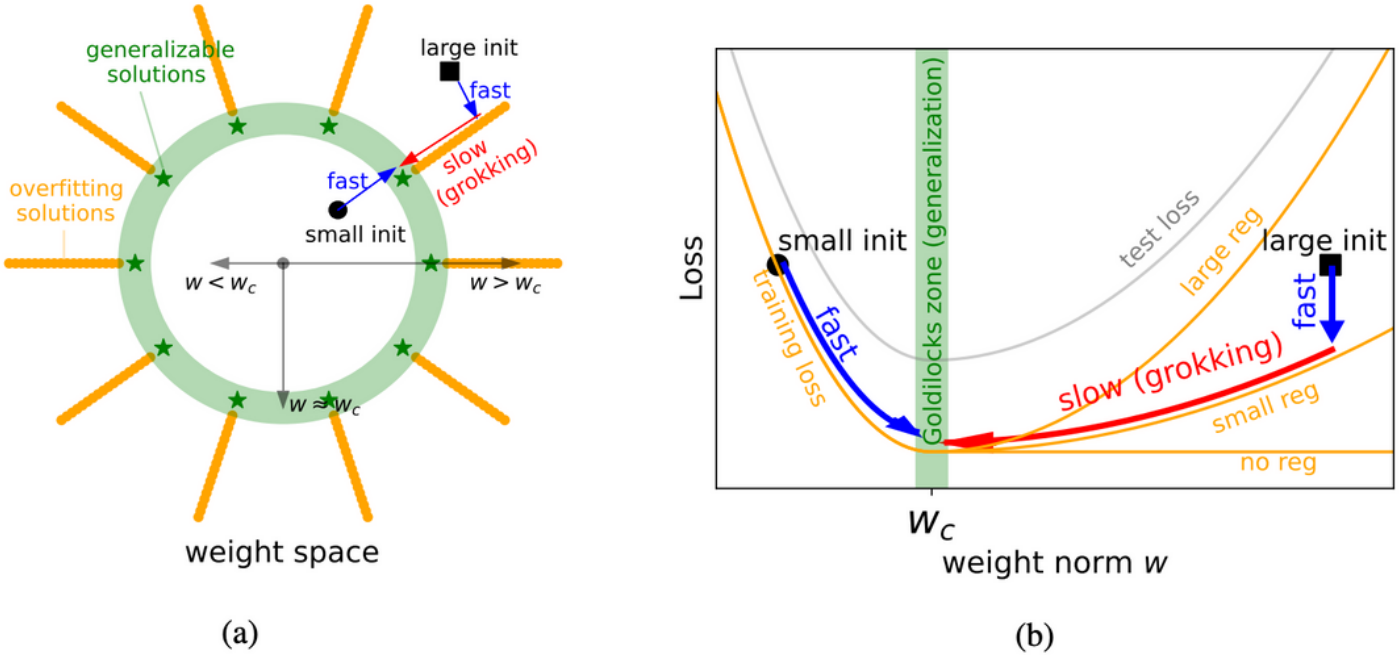
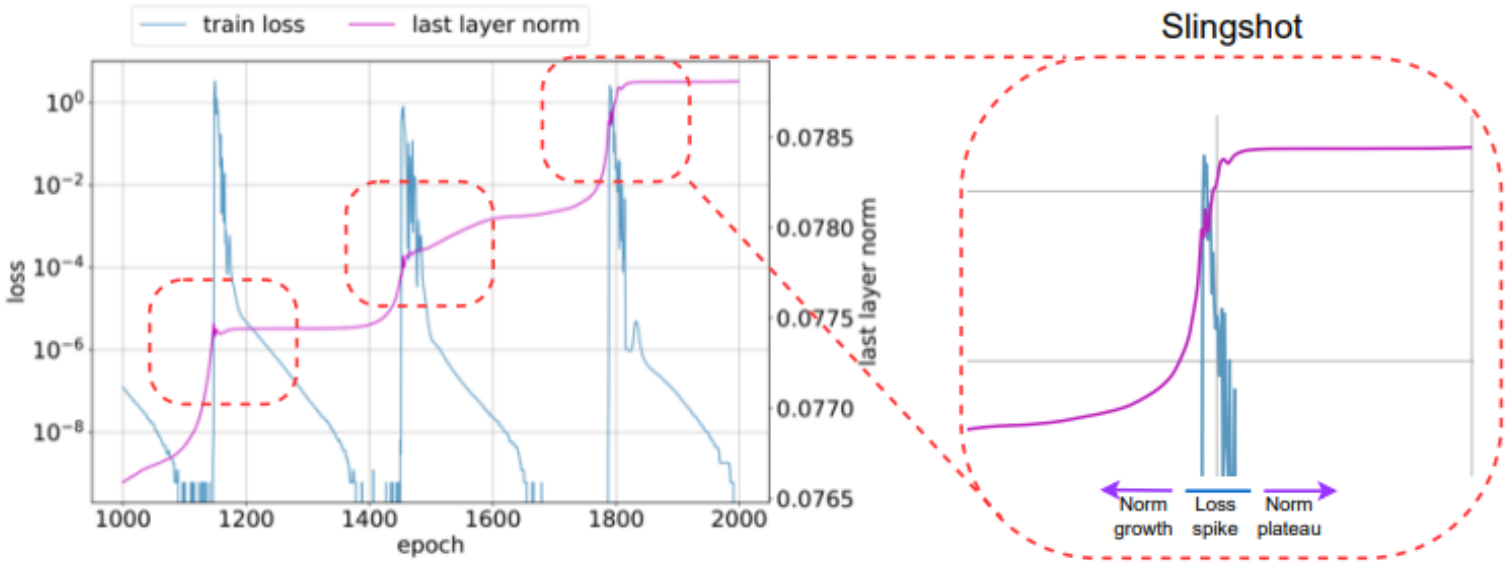


Figure 1: (a) w: L_2 norm of model weights. Generalizing solutions (green stars) are concentrated around a sphere in the weight space where $w \approx w_c$ (green). Overfitting solutions (orange) populate the $w \gtrsim w_c$ region. (b) The training loss (orange) and test loss (gray) have the shape of L and U, respectively. Their mismatch in the $w > w_c$ region leads to fast-slow dynamics, resulting in grokking. [3]

[3] Liu, Ziming, Eric J. Michaud, and Max Tegmark. "Omnigrok: Grokking beyond algorithmic data." The Eleventh International Conference on Learning Representations. 2022.

Explanation : Slingshot mechanism

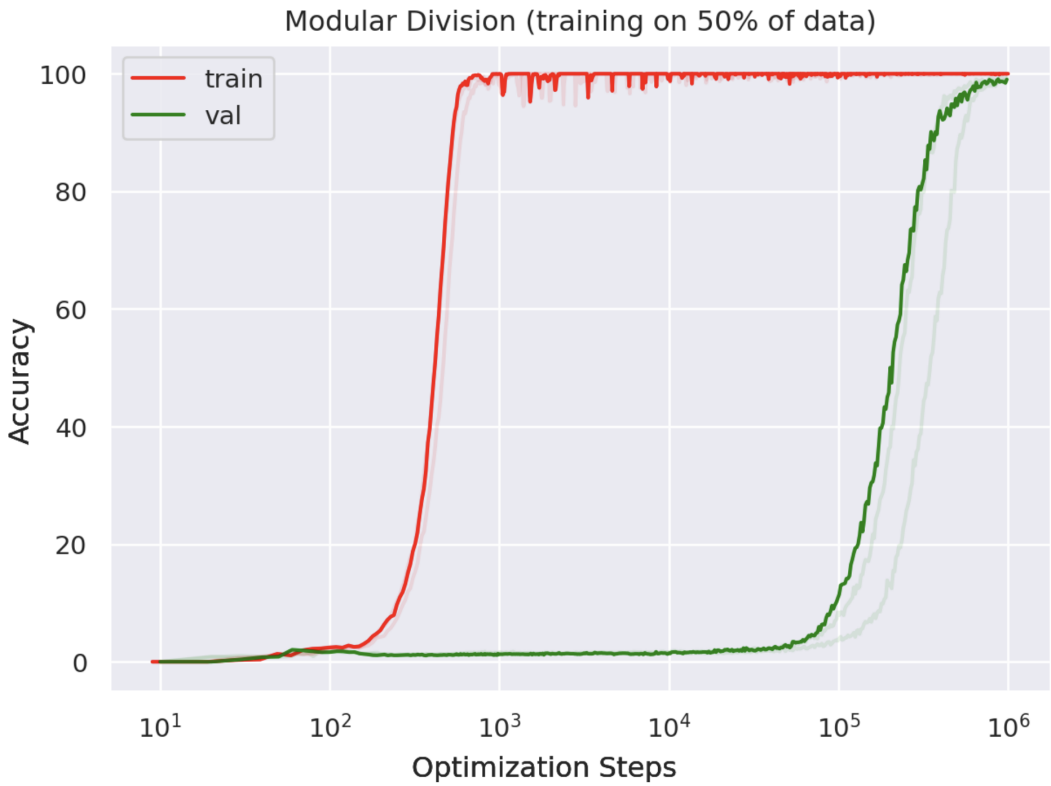
- Unstable training of Adam optimizer is key.
- $\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}}$



4

[4] Thilak, Vimal, et al. "The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon." arXiv preprint arXiv:2206.04817 (2022).

Understanding Grokking in modular arithmetic



- Condition 1: delayed rising of test accuracy.
- Condition 2: Perfect test accuracy ($> 99\%$).

Understanding Grokking in modular arithmetic

Motivation

Target task: modular arithmetic

Setup:

- $\mathbb{Z}_p := \{0, 1, \ldots, p-1\}$.
- Rule of modular arithmetic:

Binary operation

$$x_1 + x_2 = z, \quad \text{if} \quad x_1 + x_2 = z \pmod{p},$$
$$x_1 \times x_2 = z, \quad \text{if} \quad x_1 \times x_2 = z \pmod{p},$$

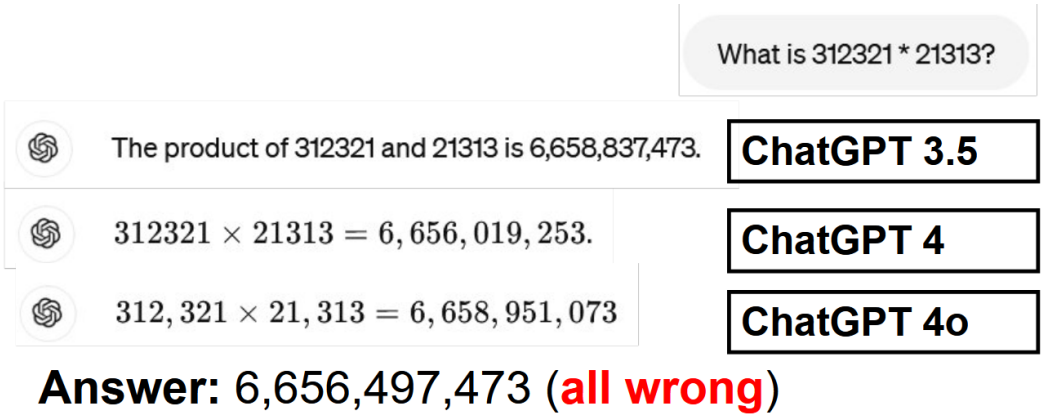
n-ary operation

$$x_1 + x_2 + x_3 = z, \quad x_1 \times x_2 \times x_3 = z.$$

A system of equations with unknowns

$$\begin{cases} 2 + A = 5 \\ A + 4 = B \end{cases} \text{ ask } B \text{ and } \begin{cases} 1 + 2 = A \\ A + 3 = B \end{cases} \text{ ask } A.$$

Motivation: Failure of GPT

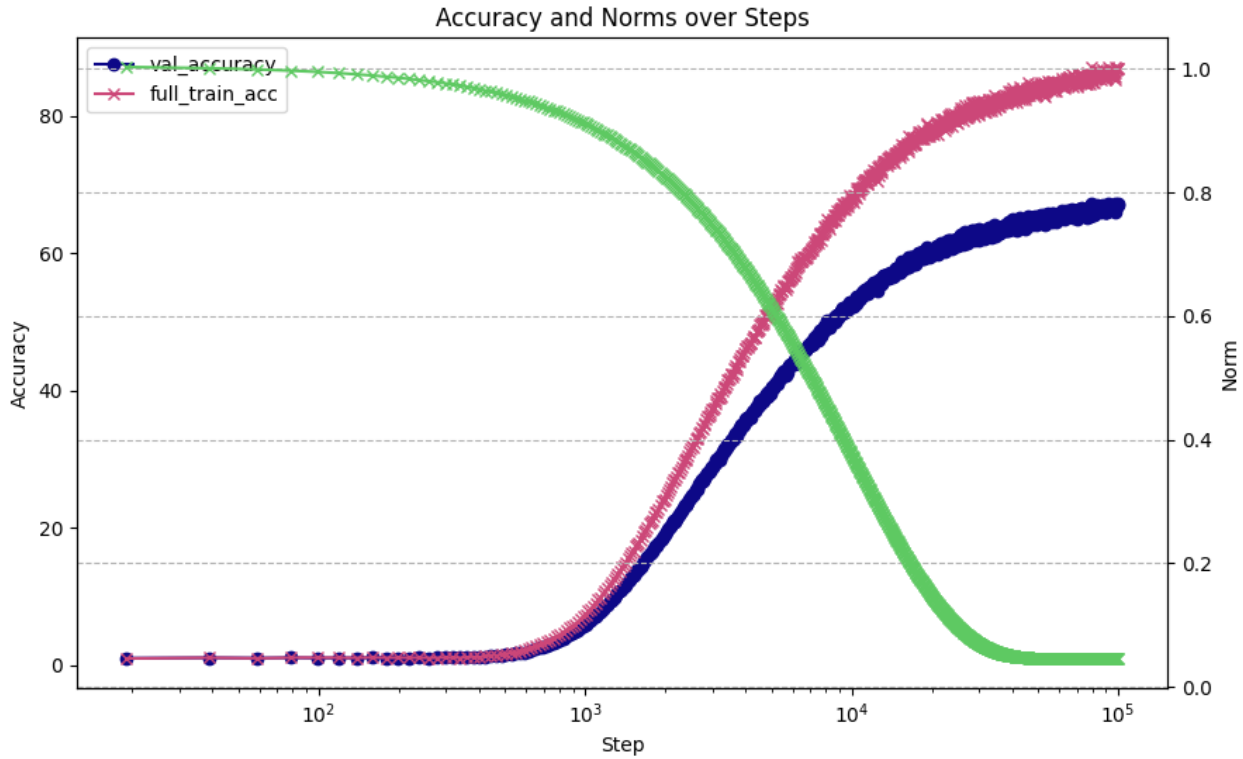


Answer: 6,656,497,473 (**all wrong**)

Motivation: Failure on difficult arithmetic

Objective:

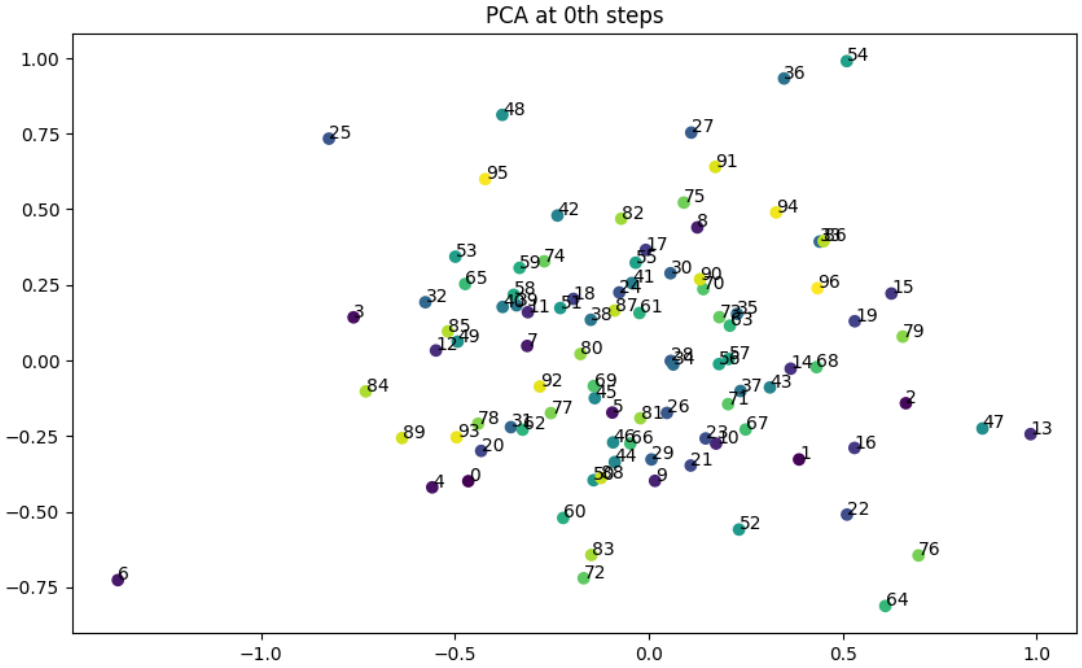
$$x_1 + x_2 + x_3 = z \tag{2}$$

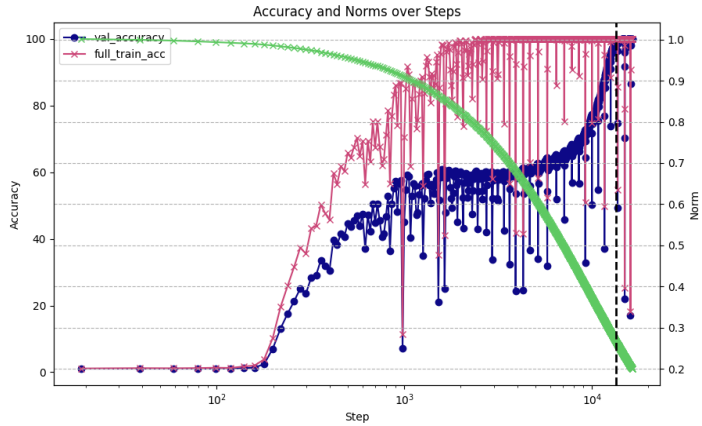


Motivation : Learning Intrinsic Feature

Objective:

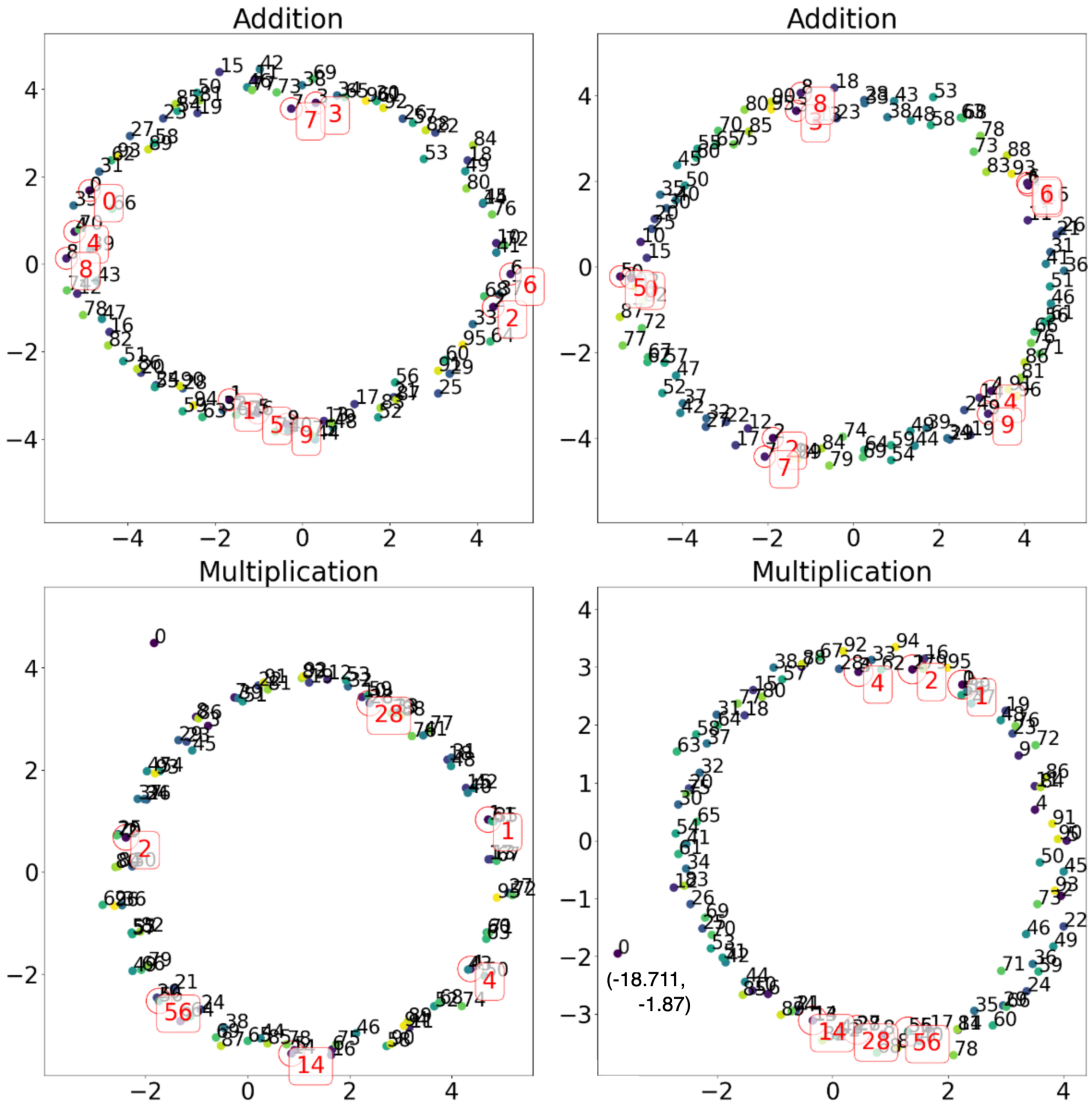
$$x_1 + x_2 = z \tag{3}$$





PCA visualization of tokens and group structure

For $p = 97$, additive group and multiplicative group has an isomorphism $\mathbb{Z}_{97} \cong C_{97}$, $\mathbb{Z}_{97}^* \cong C_{96}$ with generators $< 1 >, < 14 >$.



Kolmogorov-Arnold representation and modular arithmetic

Kolmogorov-Arnold representation

- Hilbert's 13th problem:

$$x^7 + ax^3 + bx^2 + cx + 1 = 0, \tag{4}$$

$x(a, b, c)$ can be expressed as the composition of only two variables?

- Kolmogorov and Arnold solve this problem in general form.

Theorem

Kolmogorov-Arnold representation theorem Let f be a multivariate continuous function. Then f can be represented by sinlge-variable functions and addition:

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \Psi_q \left(\sum_{i=1}^{n} \phi_{q,i}(x_i) \right) \tag{5}$$

- In 1956, Kolmogorov: three variables for any continuous function
- In 1957, Arnold: two variables
- In 1957: Kolmogorov: only one variables.

Kolmogorov-Arnold representation on binary operation

Theorem (Variants of Kolmogorov-Arnold representation)

For a closed interval $\mathcal{K}$ and any continuous function $f_n : \mathcal{K} \to \mathbb{R}$, there exist

$$\phi_{f_n} : \mathcal{K}^n \to \mathbb{R}^{2n+1} \quad \textit{and} \quad \psi_{f_n} : \mathbb{R}^{2n+1} \to \mathbb{R},$$

$$f_n(x_1, \ldots, x_n) = \psi_{f_n}\left(\sum_{i=1}^{n} \lambda_i \phi_{f_n}(x_i)\right),$$

which indicates that the embedding dimension is $2n+1$. In particular, for any binary operation $\circ$, there exist

$$\phi : \mathcal{K} \to \mathbb{R}^5 \quad \textit{and} \quad \psi : \mathbb{R}^5 \to \mathbb{R},$$

$$x_1 \circ x_2 = \psi\left(\sum_{i=1}^{2} \lambda_i \phi(x_i)\right),$$

which indicates that the embedding dimension is 5.

Permutation-invariant representation [5]

Theorem (permutation-invariant representation)

For a closed interval $\mathcal{K}$ and any permutation-invariant continuous function $f_n : \mathcal{K}^n \to \mathbb{R}$, there exist

$$\phi_{f_n} : \mathcal{K} \to \mathbb{R}^{n+1}, \quad \psi_{f_n} : \mathbb{R}^{n+1} \to \mathbb{R}, \quad f_n(x_1, \ldots, x_n) = \psi_{f_n}\left(\sum_{i=1}^{n} \phi_{f_n}(x_i)\right),$$

with embedding dimension of $n+1$. In particular, for any permutation-invariant binary operation $\circ$, we have

$$\phi_{<G,\circ>} : G \to \mathbb{R}^3 \quad \textit{and} \quad \psi_{<G,\circ>} : \mathbb{R}^3 \to G$$

with the embedding dimension of 3

$$x_1 \circ x_2 = \psi_{<G,\circ>}\left(\sum_{i=1}^{2} \phi_{<G,\circ>}(x_i)\right).$$

[5] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, A. J. Smola, Deep sets, Advances in Neural Information Processing Sys- tems 30 (2017)

Cyclic group representation

- modular addition ($\mathbb{Z}_{97} \cong C_{97}$), multiplication ($\mathbb{Z}_{97}^* \cong C_{96}$) : cyclic group

Theorem (cyclic group representation)

If G is a finite cyclic group, then there exist

$$\rho_{<G,\circ>} : G \to \mathbb{C}^* \quad \textit{and} \quad \psi_{<G,\circ>} : \mathbb{C}^* \to G$$

such that

$$\begin{aligned} x_1 \circ \cdots \circ x_n &= \psi_{<G,\circ>}\left(\sum_{i=1}^{n} \log(\rho_{<G,\circ>}(x_i))\right) \\ &= \psi_{<G,\circ>}\left(\sum_{i=1}^{n} \phi_{<G,\circ>}(x_i)\right), \end{aligned}$$

for any $x_1, \ldots, x_n \in G$, where $\phi_{<G,\circ>}(\cdot) := \log(\rho_{<G,\circ>}(\cdot))$ with common embedding dimension of 2 that is independent of n.

Example: modular addition

Example (modular addition)

For a prime number p, let $\circ$ be the modular addition in $\mathbb{Z}_p$ defined as

$$x_1 \circ x_2 := x_1 + x_2 \pmod{p}.$$

We can find the following KA representation with the embedding dimension of 2:

$$\rho(x) = \exp(\frac{2\pi i x}{p}) \in \mathbb{C}^* \cong \mathbb{R}^2 \setminus \{0\}, \quad x \in \mathbb{Z}_p,$$
$$\psi(z) = \frac{p}{2\pi i} T(z), \quad z \in \mathbb{C}^*,$$

where $T : \mathbb{C} \to \mathbb{C}$ is defined as $T(a+bi) = a + \tilde{b}$ where $a, b \in \mathbb{R},\ b = \tilde{b} + 2\pi m_b$ for $\tilde{b} \in [0, 2\pi), m_b \in \mathbb{Z}$.

$$\psi(\sum_{i=1}^{n} \phi(x_i)) = \psi(\sum_{i=1}^{n} \log(\rho(x_i)))$$
$$= \frac{p}{2\pi i} T(\frac{2\pi i x_1}{p} + \frac{2\pi i x_2}{p}) = x_1 + x_2 \pmod{p}.$$

Acceleration of Grokking

Shared structures in KA representations for various operations

For an arithmetic operation $\circ$, we recall the KA representation given by

$$x_1 \circ \cdots \circ x_n = \psi_\circ\Big(\sum_{i=1}^{n} \lambda_i \phi_\circ(x_i)\Big),$$

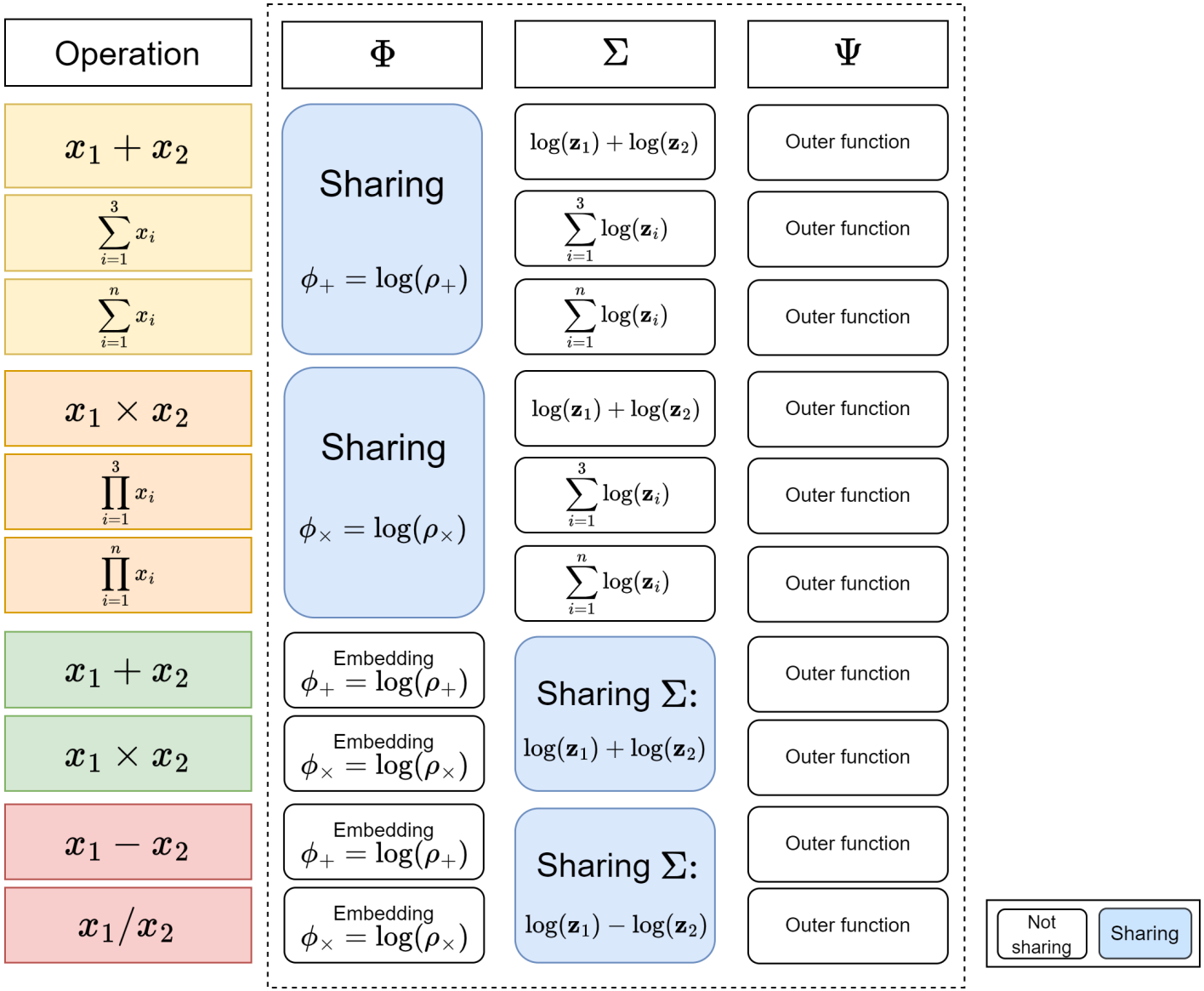
Decomposition of KA representation:

$$\begin{aligned} x_1 \circ \cdots \circ x_n &= \psi_\circ\Big(\sum_{i=1}^{n} \log \rho_\circ(x_i)\Big) \\ &= \Psi(\Sigma(\Phi(x_1, \ldots, x_n))). \end{aligned}$$

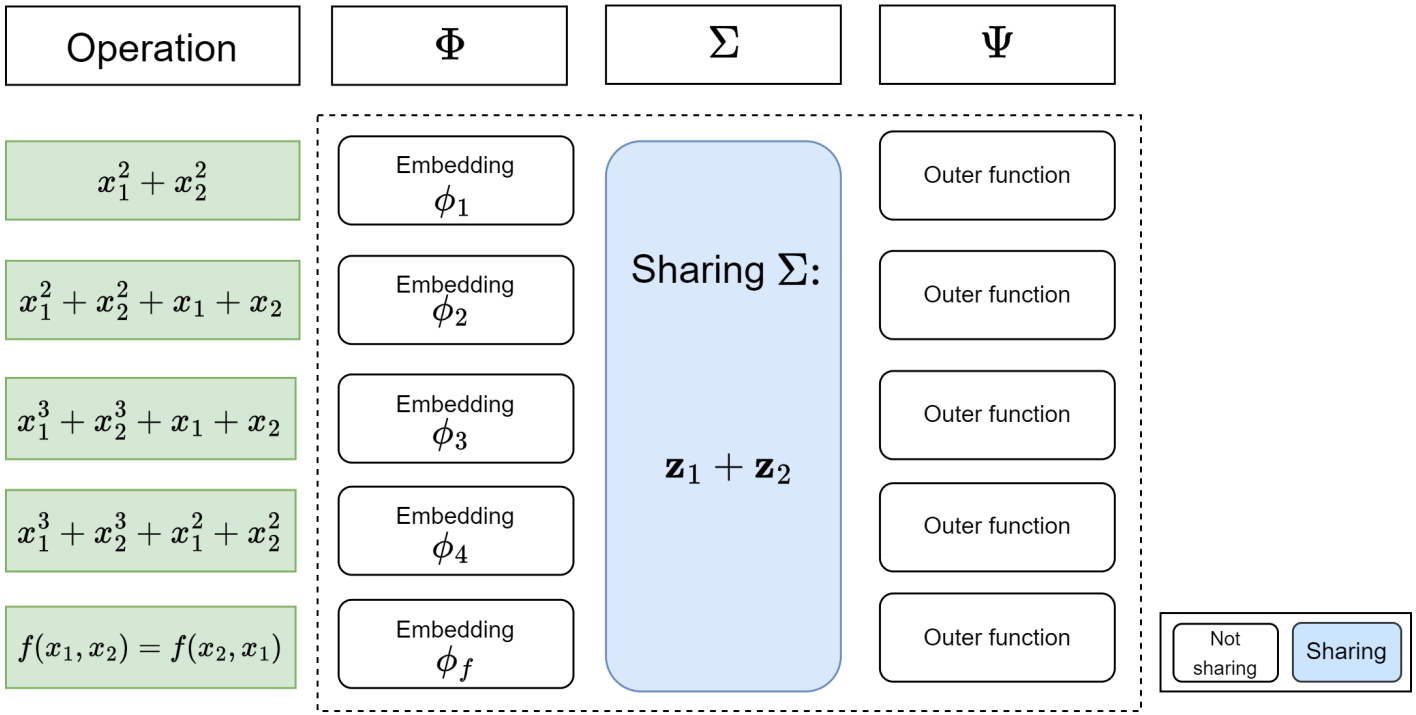
where

$$\begin{aligned} \Phi(x_1, \ldots, x_n) &= [\rho_\circ(x_1), \ldots, \rho_\circ(x_n)], \\ \Sigma(\mathbf{z}_1, \ldots, \mathbf{z}_n) &= \sum_{i=1}^{n} \log(\mathbf{z}_i), \\ \Psi(\mathbf{z}) &= \psi_\circ(\mathbf{z}) \in G \end{aligned}$$

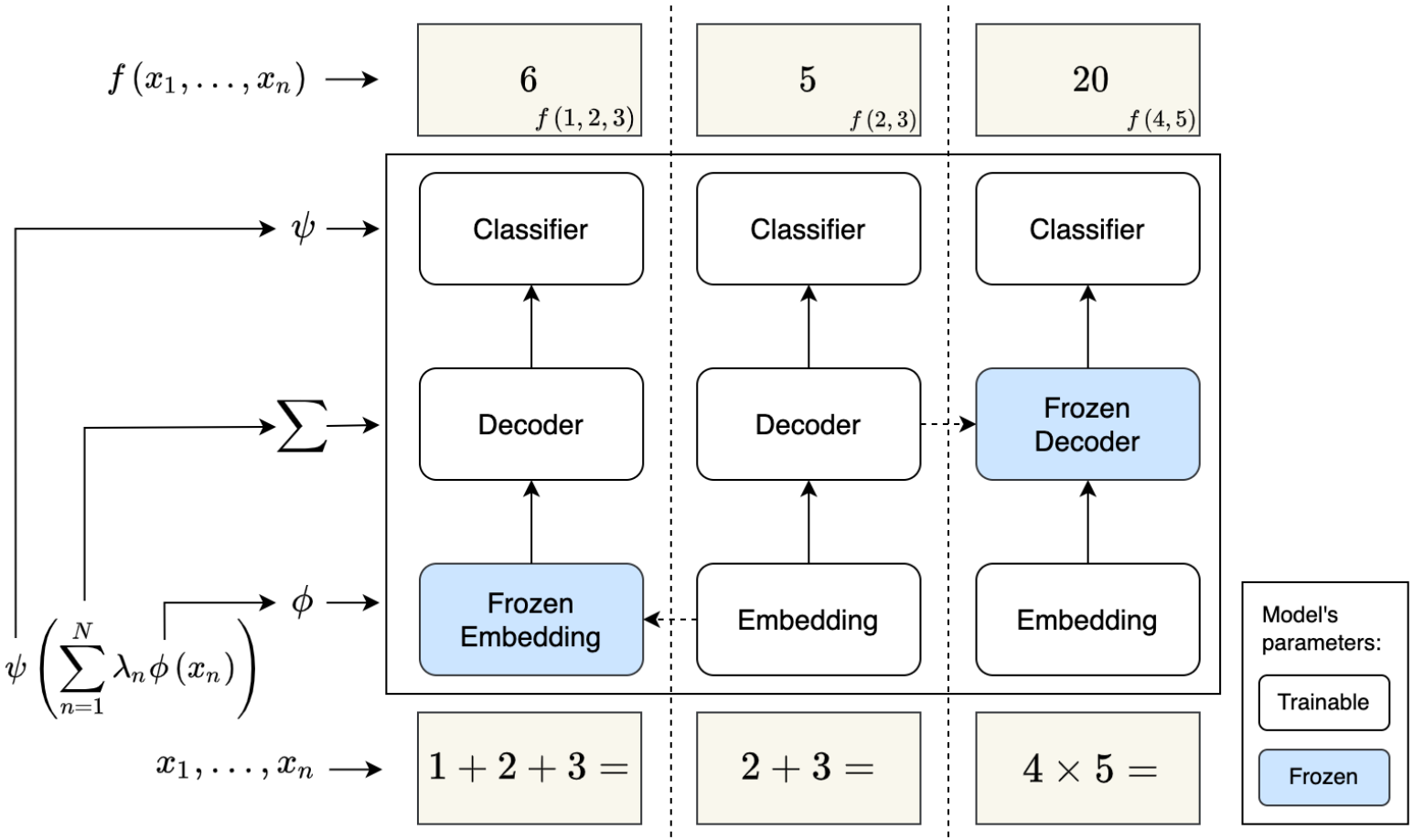
Shared structures in KA representations for various operations



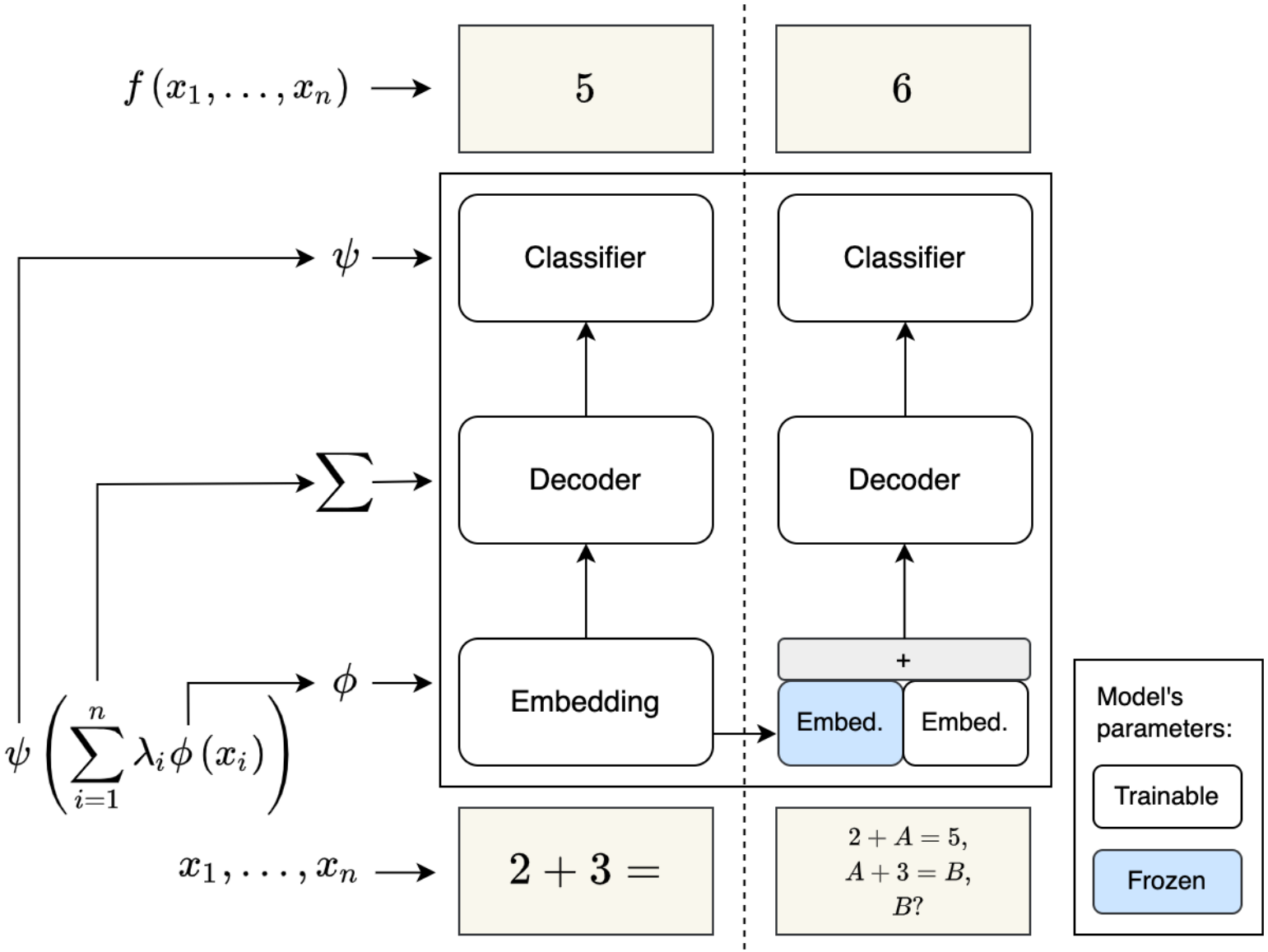
Shared structures in KA representations for various operations



Methodology : transfer learning

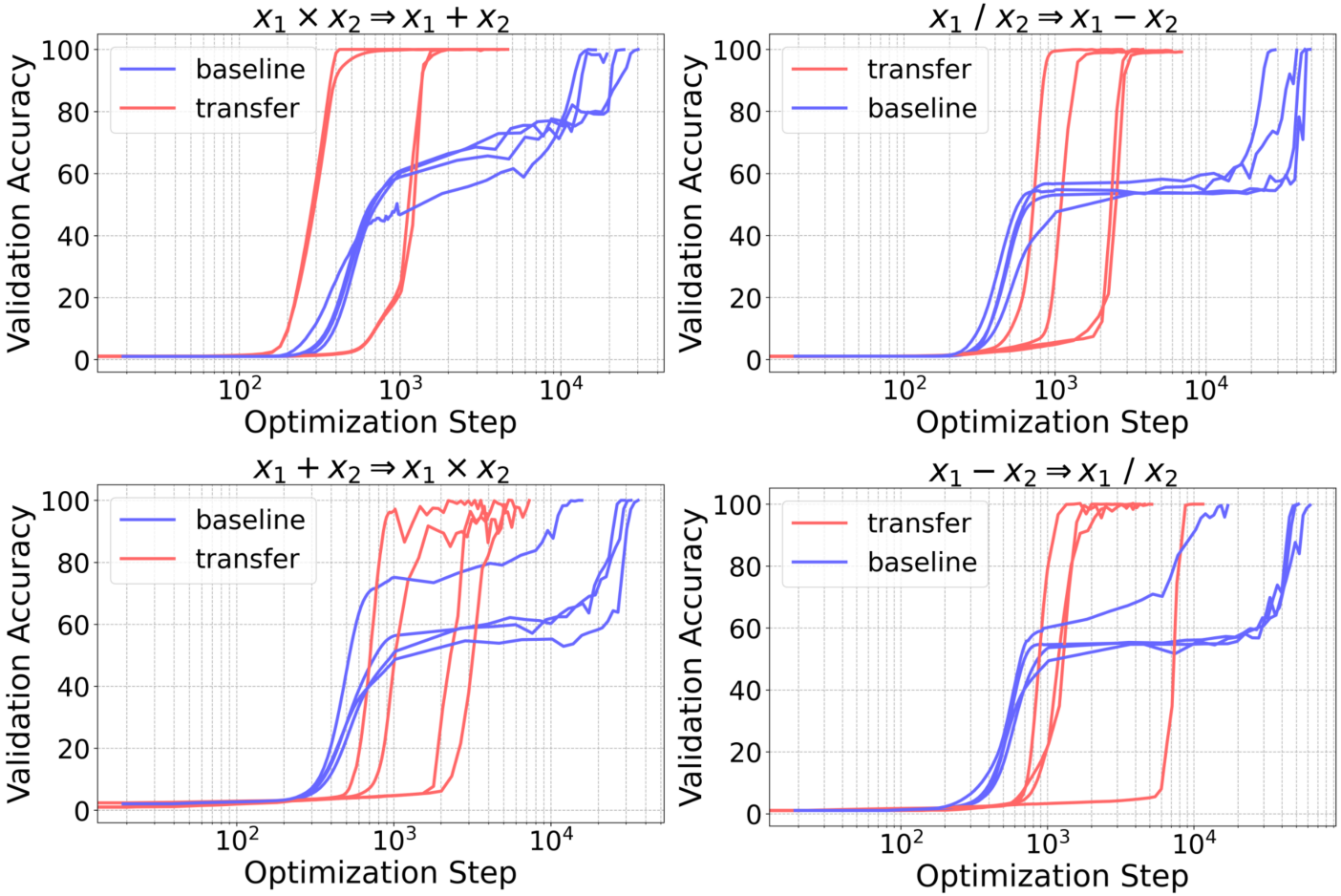


Methodology : transfer learning



Experiments

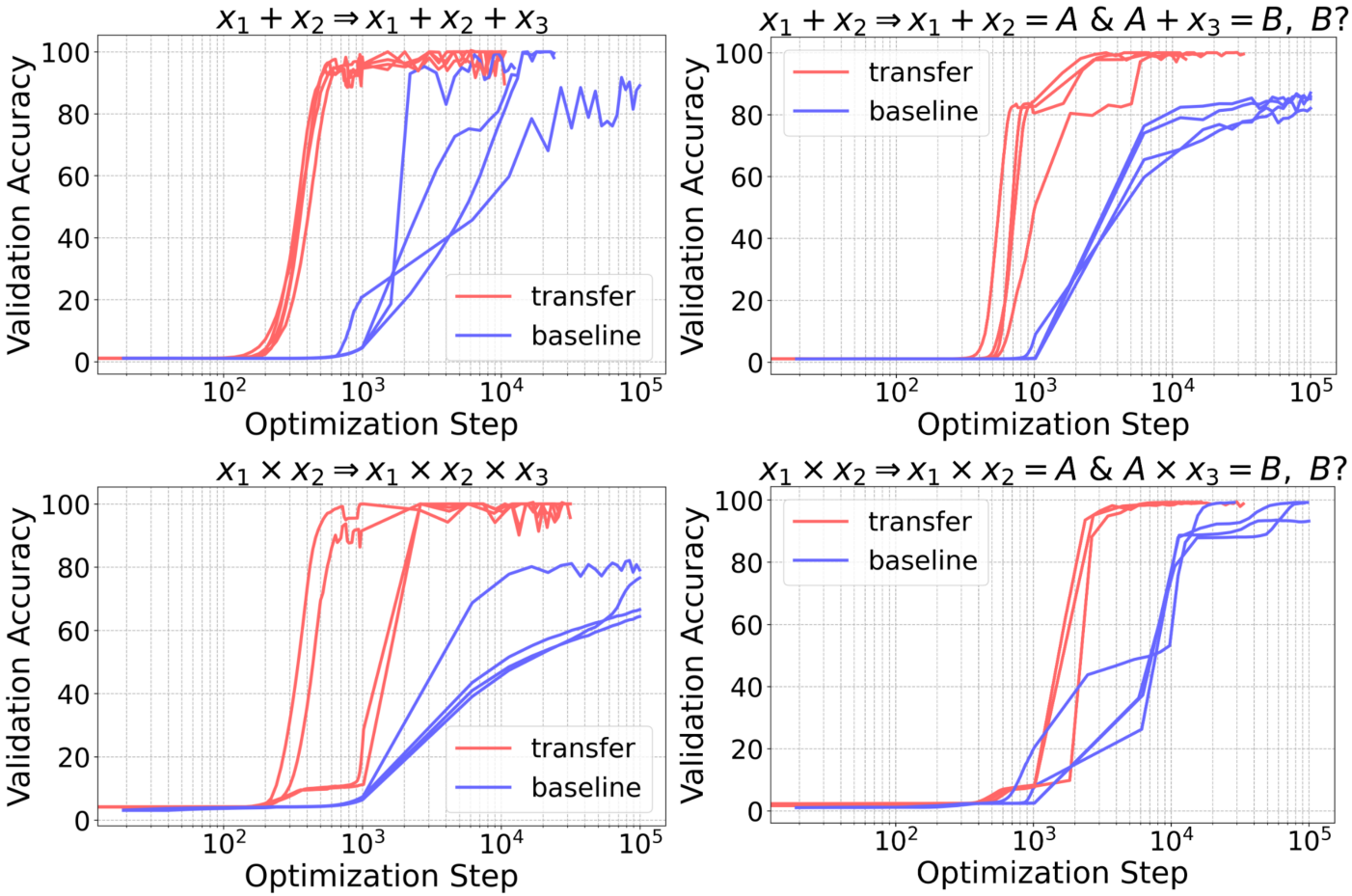
Results of decoder block transfer



Results of decoder block transfer

Operation	N	Method	Grokking step	Final accuracy
$x_1 \times x_2$	5000	baseline DT from $x+y$	20367($\pm$ 13414) 3939($\pm$ 3421)	Generalized
x_1/x_2	5000	baseline DT from $x-y$	35543($\pm$ 22517) 4115($\pm$ 3452)	Generalized
x_1+x_2	5000	baseline DT from $x_1 \times x_2$	11647($\pm$ 10705) 1087($\pm$ 642)	Generalized
x_1-x_2	5000	baseline DT from x_1/x_2	38879($\pm$ 8081) 2115($\pm$ 1288)	Generalized
$(x_1+x_2)^2$	5000	baseline DT from x_1+x_2	28459($\pm$ 9470) 2635($\pm$ 4666)	Generalized
$(x_1+x_2)^3$	5000	baseline DT from x_1+x_2	Non-generalized 479($\pm$ 107)	77.01($\pm$1.74) Generalized
${x_1}^2+{x_2}^2+x_1+x_2$	5000	baseline DT from ${x_1}^2+{x_2}^2$	755($\pm$ 293) 359($\pm$ 50)	Generalized
${x_1}^3+{x_2}^3+x_1+x_2$	5000	baseline DT from ${x_1}^2+{x_2}^2$	6831($\pm$ 4289) 331($\pm$ 36)	Generalized
${x_1}^3+{x_2}^3+{x_1}^2+{x_2}^2$	5000	baseline DT from ${x_1}^2+{x_2}^2$	11031($\pm$ 6827) 287($\pm$ 22)	Generalized
${x_1}^3+{x_2}^3+{x_1}^2+{x_2}^2+x_1+x_2$	5000	baseline DT from ${x_1}^2+{x_2}^2$	11403($\pm$ 15884) 327($\pm$ 36)	Generalized

Results of embedding transfer



Results of embedding transfer

Operation	N	Method	Grokking step	Final accuracy
$x_1 + x_2 + x_3$	10000	baseline	Non-generalized	32.17($\pm$ 11.01)
		ET	29491($\pm$ 22414)	Generalized
	100000	baseline	22186($\pm$ 25305)	74.98($\pm$ 22.92)
		ET	3039($\pm$ 5478)	Generalized
	300000	baseline	42018($\pm$ 13873)	80.70($\pm$ 19.20)
		ET	10303($\pm$ 10661)	Generalized
$x_1 \times x_2 \times x_3$	10000	baseline	Non-generalized	25.74($\pm$ 6.81)
		ET	11867($\pm$ 7080)	Generalized
	100000	baseline	Non-generalized	54.91($\pm$ 10.94)
		ET	1087($\pm$ 380)	Generalized
	300000	baseline	36852($\pm$ 46539)	78.78($\pm$ 25.56)
		ET	2591($\pm$ 407)	Generalized
$x_1 + x_2 + x_3 + x_4$	100000	baseline	Non-generalized	63.62
		ET	3559	Generalized
$x_1 \times x_2 \times x_3 \times x_4$	100000	baseline	Non-generalized	67.11
		ET	1737	Generalized

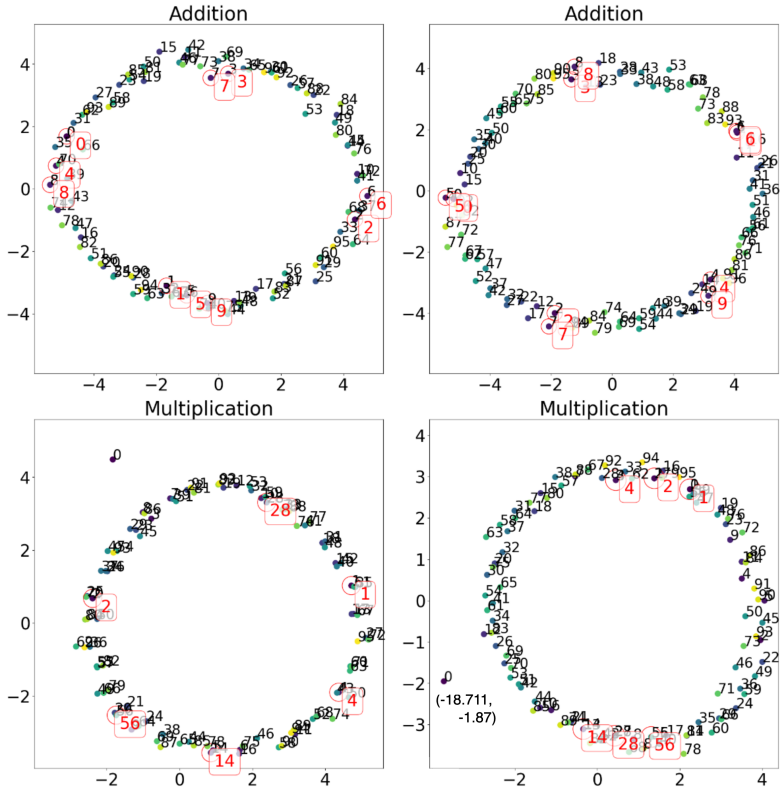
Experiments on limited token

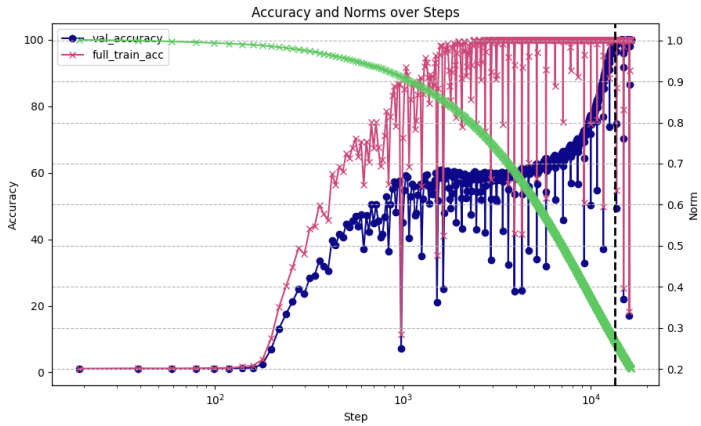
Table 1: Number of steps required for grokking when a fewer number of tokens $(80 < 97)$ are used with $N = 10,000$ training samples and trained via embedding transfer (ET).

Operation	N	Method	# of tokens	Grokking step	Final accuracy
$x_1 + x_2 + x_3$	10000	baseline	80	Non-generalized	26.42($\pm$ 0.10)
		ET	80	61034($\pm$ 25437)	Generalized
$x_1 \times x_2 \times x_3$	10000	baseline	80	Non-generalized	28.68($\pm$ 0.11)
		ET	80	10679($\pm$ 2837)	Generalized

Hypothesis

- Feature dimension for modular arithmetic : 2
- Feature learning is toward reducing feature dimension (Hypothesis).
- Feature learning induces grokking after sufficiently reducing.
- Grokking is frequent on small feature data.





Thank you