

UNRAVELING ARITHMETIC IN LARGE LANGUAGE MODELS: THE ROLE OF ALGEBRAIC STRUCTURES

Fu-Chieh Chang

MediaTek Research, Taipei, Taiwan

Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan

d09942015@ntu.edu.tw

You-Chen Lin

Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan

r13921A30@ntu.edu.tw

Pei-Yuan Wu

Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan

peiyuanwu@ntu.edu.tw

ABSTRACT

The reasoning abilities of large language models (LLMs) have improved with chain-of-thought (CoT) prompting, allowing models to solve complex tasks stepwise. However, training CoT capabilities requires detailed reasoning data, which is often scarce. The self-taught reasoner (STaR) framework addresses this by using reinforcement learning to automatically generate reasoning steps, reducing reliance on human-labeled data. Although STaR and its variants have demonstrated empirical success, a theoretical foundation explaining these improvements is lacking. Large language models (LLMs) have demonstrated remarkable mathematical capabilities, largely driven by chain-of-thought (CoT) prompting, which decomposes complex reasoning into step-by-step solutions. This approach has enabled significant advancements, as evidenced by performance on benchmarks like GSM8K and MATH. However, the mechanisms underlying LLMs’ ability to perform arithmetic in a single step of CoT remain poorly understood. Existing studies debate whether LLMs encode numerical values or rely on symbolic reasoning, while others explore attention and multi-layered processing in arithmetic tasks. In this work, we propose that LLMs learn arithmetic by capturing algebraic structures, such as commutativity and identity properties. Since these structures are observable through input-output relationships, they can generalize to unseen data. We empirically demonstrate that LLMs can learn algebraic structures using a custom dataset of arithmetic problems, as well as providing theoretical evidence showing that, under specific configurations of weights and biases, the transformer-based LLMs can generate embeddings that remain invariant to both permutations of input tokens and the presence of identity elements. Our findings indicate that leveraging algebraic structures can enhance the LLMs’ arithmetic capabilities, offering insights into improving their arithmetic performance.

1 INTRODUCTION

With the advancement of large-language models (LLMs), their mathematical capabilities have become a crucial factor in their success. This progress is largely attributed to chain-of-thought (CoT) prompting Wei et al. (2022), which enables LLMs to move beyond pattern matching and tackle complex reasoning problems through step-by-step guidance. The mathematical prowess of LLMs is demonstrated by several commercial models OpenAI (2024); Anthropic (2024) that have achieved notable success on benchmarks such as GSM8K Cobbe et al. (2021) and MATH Hendrycks et al. (2021).

Despite their achievements in solving arithmetic problems, the underlying mechanisms through which LLMs learn arithmetic operations from training data remain unclear. Although CoT explains how breaking down tasks into smaller steps facilitates mathematical problem solving, how LLMs process arithmetic tokens and execute arithmetic operations within individual steps of CoT are not yet well understood. Some studies Fangwei Zhu (2024); Levy & Geva (2024) suggest that LLMs encode numerical values, while others Deng et al. (2024) propose that LLMs learn symbolic relationships between numbers rather than directly encoding their values. Furthermore, studies such as Gorceix et al. (2024) suggest that LLMs can learn mathematical rules. However, no consensus has been reached on how or why LLMs are capable of arithmetic reasoning. A detailed literature review is shown in Sec.A.1.

In this work, we provide a novel point of view on how LLMs acquire arithmetic abilities, suggesting that such skills stem from the learning of algebraic structures such as commutativity and identity. Since LLMs only observe input and output tokens rather than explicit numerical values, direct token-to-number mapping is challenging. Instead, LLMs infer algebraic structures by examining the relationship between inputs and outputs. Previous work Karjol et al. (2023; 2024); Yang et al. (2023) has demonstrated that machine learning methods can learn symmetric structures from training data; however, these studies have not yet connected such findings to arithmetic learning in LLMs. In this work, we present both empirical and theoretical support showing that LLMs can learn algebraic structures from training data, generalizing these structures to unseen inputs. Our main contributions are as follows.

- **Empirical Evidence:** By constructing a dataset of arithmetic problems and splitting it into training and testing sets, we demonstrate that LLMs can learn and generalize the *commutativity* and *identity* properties to unseen inputs.
- **Theoretical Construction:** We provide a constructive proof illustrating how transformer-based models, with specific weight and bias configurations, can preserve hidden-state invariance under token permutations and the insertion of identity elements.

Overall, these findings suggest that LLMs can internalize algebraic structures, providing a foundation for designing strategies to further enhance their arithmetic capabilities.

2 METHODOLOGY

2.1 PROBLEM SETTINGS

We demonstrate that LLMs can learn algebraic structures from training samples and generalize to previously unseen instances. To keep our focus clear, we concentrate on arithmetic problems represented by numeric and operator symbols, rather than natural language or real-world contexts. Moreover, our study is set within a finite Abelian group (see Sec.A.1). A well-known example of a finite Abelian group is \mathbb{Z}_n , the set of integers modulo n under addition modulo n . For example, in \mathbb{Z}_5 , the elements are $\{0, 1, 2, 3, 4\}$, and the group operation is defined as $a + b \bmod 5$. The identity element of this group is 0, and every element $a \in \mathbb{Z}_n$ has an inverse $b \in \mathbb{Z}_n$ such that $a + b \equiv 0 \bmod 5$. In this work, we analyze the group operation properties of *commutativity* and *identity* in \mathbb{Z}_n .

2.2 DATASET FOR COMMUTATIVITY AND IDENTITY

In this work, we show that LLMs can acquire the concepts of commutativity and identity purely from the dataset we provide, rather than relying on any preexisting, pre-trained knowledge. To validate this, we construct a dataset of addition problems in \mathbb{Z}_n . Each element in \mathbb{Z}_n is denoted as z_i , where $z_0 = 0, z_1 = 1, \dots, z_{n-1} = n - 1$. The dataset consists of addition problems with M terms, formally expressed as:

$$z_{i_1} + z_{i_2} + \dots + z_{i_M} = z_{(i_1+i_2+\dots+i_M) \bmod n}, \quad (1)$$

where $0 \leq i_1, i_2, \dots, i_M < n$. Here, mod denotes the modulo operator. For each problem, “ $z_{i_1} + z_{i_2} + \dots + z_{i_M} =$ ” serves as input tokens, while the label is “ $z_{(i_1+i_2+\dots+i_M) \bmod n}$ ”. Thus, the LLM must predict the correct element “ $z_{(i_1+i_2+\dots+i_M) \bmod n}$ ” given the inputs. In addition, we

Training, for Operator “+”		Testing, for Operator “+”	
Commutativity	Identity	Commutativity	Identity
$z_3 + z_4 + z_5 + z_5 + z_5 + z_6 = z_0$	$z_0 + z_4 + z_3 + z_5 + z_3 + z_1 = z_2$	$z_2 + z_3 + z_3 + z_5 + z_5 + z_6 = z_3$	$z_0 + z_5 + z_2 + z_5 + z_2 + z_3 = z_3$
$z_4 + z_5 + z_3 + z_5 + z_5 + z_6 = z_0$	$z_4 + z_0 + z_3 + z_5 + z_3 + z_1 = z_2$	$z_3 + z_3 + z_5 + z_5 + z_6 + z_2 = z_3$	$z_5 + z_0 + z_2 + z_5 + z_2 + z_3 = z_3$
$z_3 + z_5 + z_5 + z_6 + z_4 + z_5 = z_0$	$z_4 + z_3 + z_0 + z_5 + z_3 + z_1 = z_2$	$z_5 + z_6 + z_5 + z_3 + z_2 + z_3 = z_3$	$z_5 + z_2 + z_0 + z_5 + z_2 + z_3 = z_3$
$z_4 + z_5 + z_5 + z_6 + z_3 + z_5 = z_0$	$z_4 + z_3 + z_5 + z_0 + z_3 + z_1 = z_2$	$z_3 + z_2 + z_5 + z_3 + z_6 + z_5 = z_3$	$z_5 + z_2 + z_5 + z_0 + z_2 + z_3 = z_3$
$z_6 + z_5 + z_5 + z_3 + z_5 + z_4 = z_0$	$z_4 + z_3 + z_5 + z_3 + z_0 + z_1 = z_2$	$z_5 + z_3 + z_5 + z_3 + z_2 + z_6 = z_3$	$z_5 + z_2 + z_5 + z_2 + z_0 + z_3 = z_3$
$z_3 + z_4 + z_5 + z_5 + z_6 + z_5 = z_0$	$z_4 + z_3 + z_5 + z_3 + z_1 + z_0 = z_2$	$z_6 + z_3 + z_5 + z_3 + z_2 + z_5 = z_3$	$z_5 + z_2 + z_5 + z_2 + z_3 + z_0 = z_3$
$z_3 + z_6 + z_4 + z_5 + z_5 + z_5 = z_0$	$z_4 + z_3 + z_5 + z_3 + z_1 = z_2$		
$z_6 + z_3 + z_5 + z_3 + z_5 + z_2 = z_3$	$z_5 + z_2 + z_5 + z_2 + z_3 = z_3$		
Training, for Operator \oplus		Testing, for Operator \oplus	
Commutativity	Identity	Commutativity	Identity
$z_3 \oplus z_4 \oplus z_5 \oplus z_5 \oplus z_5 \oplus z_6 = r_2$	$z_0 \oplus z_4 \oplus z_3 \oplus z_5 \oplus z_3 \oplus z_1 = r_1$	$z_2 \oplus z_3 \oplus z_3 \oplus z_5 \oplus z_5 \oplus z_6 = r_4$	$z_0 \oplus z_5 \oplus z_2 \oplus z_5 \oplus z_2 \oplus z_3 = r_5$
$z_4 \oplus z_5 \oplus z_3 \oplus z_5 \oplus z_5 \oplus z_6 = r_2$	$z_4 \oplus z_0 \oplus z_3 \oplus z_5 \oplus z_3 \oplus z_1 = r_1$	$z_3 \oplus z_3 \oplus z_5 \oplus z_5 \oplus z_6 \oplus z_2 = r_4$	$z_5 \oplus z_0 \oplus z_2 \oplus z_5 \oplus z_2 \oplus z_3 = r_5$
$z_3 \oplus z_5 \oplus z_5 \oplus z_6 \oplus z_4 \oplus z_5 = r_2$	$z_4 \oplus z_3 \oplus z_0 \oplus z_5 \oplus z_3 \oplus z_1 = r_1$	$z_5 \oplus z_6 \oplus z_5 \oplus z_3 \oplus z_2 \oplus z_3 = r_4$	$z_5 \oplus z_2 \oplus z_0 \oplus z_5 \oplus z_2 \oplus z_3 = r_5$
$z_4 \oplus z_5 \oplus z_5 \oplus z_6 \oplus z_3 \oplus z_5 = r_2$	$z_4 \oplus z_3 \oplus z_5 \oplus z_0 \oplus z_3 \oplus z_1 = r_1$	$z_3 \oplus z_2 \oplus z_5 \oplus z_3 \oplus z_6 \oplus z_5 = r_4$	$z_5 \oplus z_2 \oplus z_5 \oplus z_0 \oplus z_2 \oplus z_3 = r_5$
$z_6 \oplus z_5 \oplus z_5 \oplus z_3 \oplus z_5 \oplus z_4 = r_2$	$z_4 \oplus z_3 \oplus z_5 \oplus z_3 \oplus z_0 \oplus z_1 = r_1$	$z_5 \oplus z_3 \oplus z_5 \oplus z_3 \oplus z_2 \oplus z_6 = r_4$	$z_5 \oplus z_2 \oplus z_5 \oplus z_2 \oplus z_0 \oplus z_3 = r_5$
$z_3 \oplus z_4 \oplus z_5 \oplus z_5 \oplus z_6 \oplus z_5 = r_2$	$z_4 \oplus z_3 \oplus z_5 \oplus z_3 \oplus z_1 \oplus z_0 = r_1$	$z_6 \oplus z_3 \oplus z_5 \oplus z_3 \oplus z_2 \oplus z_5 = r_4$	$z_5 \oplus z_2 \oplus z_5 \oplus z_2 \oplus z_3 \oplus z_0 = r_5$
$z_3 \oplus z_6 \oplus z_4 \oplus z_5 \oplus z_5 \oplus z_5 = r_2$	$z_4 \oplus z_3 \oplus z_5 \oplus z_3 \oplus z_1 = r_1$		
$z_6 \oplus z_3 \oplus z_5 \oplus z_3 \oplus z_5 \oplus z_2 = r_4$	$z_5 \oplus z_2 \oplus z_5 \oplus z_2 \oplus z_3 = r_5$		
Training, for Operator \ominus		Testing, for Operator \ominus	
$z_3 \ominus z_4 \ominus z_5 \ominus z_5 \ominus z_5 \ominus z_6 = 2$	$z_0 \ominus z_4 \ominus z_3 \ominus z_5 \ominus z_3 \ominus z_1 = 3$	$z_2 \ominus z_3 \ominus z_3 \ominus z_5 \ominus z_5 \ominus z_6 = 2$	$z_0 \ominus z_5 \ominus z_2 \ominus z_5 \ominus z_2 \ominus z_3 = 2$
$z_4 \ominus z_5 \ominus z_3 \ominus z_5 \ominus z_5 \ominus z_6 = 2$	$z_4 \ominus z_0 \ominus z_3 \ominus z_5 \ominus z_3 \ominus z_1 = 3$	$z_3 \ominus z_3 \ominus z_5 \ominus z_5 \ominus z_6 \ominus z_2 = 3$	$z_5 \ominus z_0 \ominus z_2 \ominus z_5 \ominus z_2 \ominus z_3 = 2$
$z_3 \ominus z_5 \ominus z_5 \ominus z_6 \ominus z_4 \ominus z_5 = 2$	$z_4 \ominus z_3 \ominus z_0 \ominus z_5 \ominus z_3 \ominus z_1 = 4$	$z_5 \ominus z_6 \ominus z_5 \ominus z_3 \ominus z_2 \ominus z_3 = 3$	$z_5 \ominus z_2 \ominus z_0 \ominus z_5 \ominus z_2 \ominus z_3 = 3$
$z_4 \ominus z_5 \ominus z_5 \ominus z_6 \ominus z_3 \ominus z_5 = 2$	$z_4 \ominus z_3 \ominus z_5 \ominus z_0 \ominus z_3 \ominus z_1 = 3$	$z_3 \ominus z_2 \ominus z_5 \ominus z_3 \ominus z_6 \ominus z_5 = 3$	$z_5 \ominus z_2 \ominus z_5 \ominus z_0 \ominus z_2 \ominus z_3 = 2$
$z_6 \ominus z_5 \ominus z_5 \ominus z_3 \ominus z_5 \ominus z_4 = 4$	$z_4 \ominus z_3 \ominus z_5 \ominus z_3 \ominus z_0 \ominus z_1 = 3$	$z_5 \ominus z_3 \ominus z_5 \ominus z_3 \ominus z_2 \ominus z_6 = 3$	$z_5 \ominus z_2 \ominus z_5 \ominus z_2 \ominus z_0 \ominus z_3 = 3$
$z_3 \ominus z_4 \ominus z_5 \ominus z_6 \ominus z_5 \ominus z_5 = 2$	$z_4 \ominus z_3 \ominus z_5 \ominus z_3 \ominus z_1 \ominus z_0 = 4$	$z_6 \ominus z_3 \ominus z_5 \ominus z_3 \ominus z_2 \ominus z_5 = 3$	$z_5 \ominus z_2 \ominus z_5 \ominus z_2 \ominus z_3 \ominus z_0 = 3$
$z_3 \ominus z_6 \ominus z_4 \ominus z_5 \ominus z_5 \ominus z_5 = 3$	$z_4 \ominus z_3 \ominus z_5 \ominus z_3 \ominus z_1 = 3$		
$z_6 \ominus z_3 \ominus z_5 \ominus z_3 \ominus z_5 \ominus z_2 = 3$	$z_5 \ominus z_2 \ominus z_5 \ominus z_2 \ominus z_3 = 2$		

Figure 1: Illustration of dataset for operator “+”, \oplus and \ominus . Notice that the same set of tokens is maintained across all operators to ensure that certain token combinations appear exclusively either in the training set or the testing set, as required.

focus on the scenario that the model must directly predict the label from the input without performing any intermediate CoT reasoning. Fig. 1 and Appendix A.4 provide examples of both the training and testing sets for \mathbb{Z}_7 . In the subsequent sections, we explain how the datasets are constructed so as to test whether LLMs genuinely learn and generalize the underlying algebraic properties of commutativity and identity.

Commutativity: To demonstrate that LLMs can learn the commutative property, we begin by selecting a sequence $(z_{i_1}, z_{i_2}, \dots, z_{i_M})$ from \mathbb{Z}_n , where $z_{i_1} \leq z_{i_2} \leq \dots \leq z_{i_M}$ and $z_{i_1} > 0$, and generate sequences in the form of Eq. equation 1. Next, we sample several permutations of $(z_{i_1}, z_{i_2}, \dots, z_{i_M})$, denoting each permutation as $(z_{j_1}, z_{j_2}, \dots, z_{j_M})$, and generate sequences as in Eq. equation 1 accordingly. For any given sequence $(z_{i_1}, z_{i_2}, \dots, z_{i_M})$, to allow the model to recognize the element $z_{(i_1+i_2+\dots+i_M) \bmod n}$ in the first place, at least one permutation of each

sequence in testing set appears in the training set. However, if more than one permutations are included in the training set, all other permutations of this same sequence are excluded from the training set and placed in the testing set. This ensures that the model must generalize the commutativity property, i.e. it must infer correct outputs for unseen permutations in the testing set, based on the only one permutation learned during training. As shown in the upper row of Fig. 1, every permutation of $(z_3, z_4, z_5, z_5, z_5, z_6)$ is included in the training set (highlighted in red). Meanwhile, the testing set contains a different permutation sequences of Eq. equation 1, $(z_2, z_3, z_3, z_5, z_5, z_6)$, highlighted in orange. However, one permutation of that sequence— $(z_6, z_3, z_5, z_3, z_5, z_2)$ —does appear in the training set, ensuring that the model is exposed to at least one variant of the same element sum.

Identity: In \mathbb{Z}_n , the identity element for addition is z_0 . To verify whether LLMs learn this identity property, we first pick $M - 1$ variables, $(z_{i_1}, \dots, z_{i_{M-1}})$ from \mathbb{Z}_n , and insert z_0 among them in all

possible positions. Substituting each arrangement into Eq. equation 1 yields M distinct equations:

$$\begin{aligned} z_0 + z_{i_1} + z_{i_2} + \cdots + z_{i_{M-1}} &= z_{(i_1+i_2+\cdots+i_{M-1}) \bmod n}, \\ z_{i_1} + z_0 + z_{i_2} + \cdots + z_{i_{M-1}} &= z_{(i_1+i_2+\cdots+i_{M-1}) \bmod n}, \\ &\vdots \\ z_{i_1} + z_{i_2} + \cdots + z_{i_{M-1}} + z_0 &= z_{(i_1+i_2+\cdots+i_{M-1}) \bmod n}. \end{aligned}$$

To ensure that the model does not merely exploit permutation invariance, we assign all possible insertions of z_0 in the sequence $(z_{i_1}, \dots, z_{i_{M-1}})$ exclusively to either the training set or the testing set. Moreover, because the value of $z_{(i_1+i_2+\cdots+i_{M-1}) \bmod n}$ should be established using the equation without the identity element, the following *base equation* without z_0 must be included in the training set:

$$z_{i_1} + z_{i_2} + \cdots + z_{i_{M-1}} = z_{(i_1+i_2+\cdots+i_{M-1}) \bmod n}.$$

In the first row of Fig. 1, we illustrate how identity elements appear in both the training and testing sets. The training set includes the *base equation*, $z_4 + z_3 + z_5 + z_3 + z_1 = z_2$, along with variants where z_0 is inserted into every possible position (highlighted in blue). In contrast, the equation $z_0 + z_5 + z_2 + z_5 + z_2 + z_3 = z_3$ and its variants with z_0 inserted in different positions are placed in the testing set (highlighted in cyan). However, the base equation $z_5 + z_2 + z_5 + z_2 + z_3 = z_3$ appears in the training set, ensuring that the model can infer $z_0 + z_5 + z_2 + z_5 + z_2 + z_3 = z_3$ when recognizing z_0 as the identity element.

2.3 DATASET TO EXCLUDE NUMERICAL CALCULATION

Training on the dataset introduced in Sec. 2.2 leaves the possibility that LLMs might exploit numerical relationships inherent in the indices. Specifically, an LLM could potentially identify the input token indices and rely on computing the modulo sum of these indices to derive results, bypassing the application of commutativity and identity principles. For example, in Fig. 1, the dataset for commutativity contains $z_6 + z_3 + z_5 + z_3 + z_5 + z_2 = z_3$ in training set and $z_2 + z_3 + z_3 + z_5 + z_5 + z_6 = z_3$ in testing set. When answering the test equation, we anticipate that LLMs are able to leverage commutativity to infer z_3 from the training equation. Nonetheless, we cannot entirely rule out the possibility that an LLM could deduce the numerical value of the index $(2, 3, 3, 5, 5, 6)$ and use the sum $(2 + 3 + 3 + 5 + 5 + 6) \bmod 7 = 3$. To prevent this, we introduce a new operator \oplus which takes the same inputs as “+” but produces outputs unrelated to the inputs in a numerical sense. Specifically, given any sequence $(z_{i_1}, z_{i_2}, \dots, z_{i_M})$ in \mathbb{Z}_n where $z_{i_1} \leq z_{i_2} \leq \cdots \leq z_{i_M}$ and $z_{i_1} > 0$, the result of applying \oplus to this sequence (or any permutation of it) is a randomly selected element r_i from the set $\{r_0, r_1, \dots, r_{n-1}\}$, whose elements lie outside \mathbb{Z}_n . Hence, \oplus is invariant under input permutations. Namely,

$$z_{i_1} \oplus z_{i_2} \oplus \cdots \oplus z_{i_M} = r_i, \quad z_{i_2} \oplus z_{i_1} \oplus \cdots \oplus z_{i_M} = r_i, \quad \dots, \quad z_{i_M} \oplus z_{i_{M-1}} \oplus \cdots \oplus z_{i_1} = r_i.$$

Besides, \oplus is invariant under insertion of the identity element z_0 at any position such that

$$z_0 \oplus z_{i_1} \oplus z_{i_2} \oplus \cdots \oplus z_{i_M} = r_i, \quad \dots, \quad z_{i_1} \oplus z_{i_2} \oplus \cdots \oplus z_{i_M} \oplus z_0 = r_i.$$

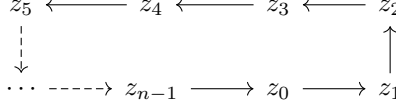
Hence, for LLMs to perform well on this dataset, they cannot rely on the numerical relationship of the index; rather, they must learn the commutative property and identity property of \oplus . We include a corresponding dataset involving \oplus in both training and testing. As shown in the second row of Fig. 1, we replace “+” with \oplus and adjust the resulting values accordingly. By comparing how “+” and \oplus perform on commutativity and identity tasks under the same training and test inputs in \mathbb{Z}_n , we can discern whether the performance of “+” benefits from numerical computation or purely from algebraic structures.

2.4 DATASET TO AVOID TRIVIAL SOLUTIONS

In this work, our goal is for LLMs to learn the commutative and identity properties specifically for the “+” and \oplus operator. If the model were trained only on the dataset described in the previous sections, it could adopt a trivial solution in which commutativity and identity trivially apply to all tokens z_i

regardless of the existence of the operator “+” or \oplus . For instance, by setting all position embeddings to zero (see Sec. 2.5). To prevent this, we include an additional dataset that features operators that lack commutativity and identity properties. Specifically, we introduce three new operators, \ominus , \triangleleft , and \triangleright . Details of these operators are provided as follows.

Operator \ominus : Counts of Encountering z_0 along the Cyclic Group \mathbb{Z}_n . It is straightforward to show that \mathbb{Z}_n is a cyclic group, as illustrated below:



We define $\ominus : \mathbb{Z}_n \times \mathbb{Z}_n \rightarrow \mathbb{N}$ as an operator that counts the number of times z_0 is encountered. Concretely, for any $z_i, z_j \in \mathbb{Z}_n$, $z_i \ominus z_j$ equals the number of occurrences of z_0 when traveling from z_i to z_j around the cyclic group.

$$z_i \rightarrow z_{(i+1) \bmod n} \rightarrow z_{(i+2) \bmod n} \rightarrow \cdots \rightarrow z_{(j-1) \bmod n} \rightarrow z_j.$$

For example, in \mathbb{Z}_5 , $z_3 \ominus z_1 = 1$ because traveling from $z_3 \rightarrow z_4 \rightarrow z_0 \rightarrow z_1$ encounters z_0 once. Similarly, $z_2 \ominus z_5 = 0$ because $z_2 \rightarrow z_3 \rightarrow z_4 \rightarrow z_5$ does not pass through z_0 . We set $z_i \ominus z_i = 1$ because one must traverse all elements of \mathbb{Z}_n to return to the same element. We also define $z_i \ominus z_j \ominus z_k$ as the total number of z_0 encounters when traveling from z_i to z_j , then continuing to z_k , such that

$$z_i \ominus z_j \ominus z_k = (z_i \ominus z_j) + (z_j \ominus z_k).$$

For instance, in \mathbb{Z}_5 , $z_4 \ominus z_2 \ominus z_1 = 2$. Besides, it is straightforward to verify that \ominus does not satisfy commutativity, nor is z_0 an identity element, because $z_i \ominus z_j \neq z_j \ominus z_i$ (for all $i \neq j$ with $i, j > 0$), and $z_i \ominus z_j \neq z_i \ominus z_0 \ominus z_j$ (for all $i < j$ with $i, j > 0$). We add a dataset involving the \ominus operator to the training set and testing set. As shown in Fig. 1, we replace “+” with \ominus and update the resulting values accordingly. Note that the output of \ominus is a natural number in \mathbb{N} rather than an element of \mathbb{Z}_n .

Operators \triangleleft and \triangleright : Left-Hand Side and Right-Hand Side Elements. Alongside the \ominus operator, we introduce two additional operators, \triangleleft and \triangleright , which also lack commutativity and identity elements. These operators simply return the left-hand-side or right-hand-side argument, respectively:

$$z_i \triangleright z_j = z_j \quad \text{and} \quad z_i \triangleleft z_j = z_i.$$

It is straightforward to see that these operators do not satisfy commutativity and that z_0 is not an identity element for either. However, they satisfy associative, so expressions such as $z_i \triangleright z_j \triangleright z_k$ produce a unique result. We add a dataset that includes these operators in both the training and testing sets. An example of the whole dataset, which includes all operators “+”, \oplus , \ominus , \triangleleft and \triangleright is shown in Sec. A.4. In the following section, we discuss how LLMs can learn the underlying algebraic structures from these datasets.

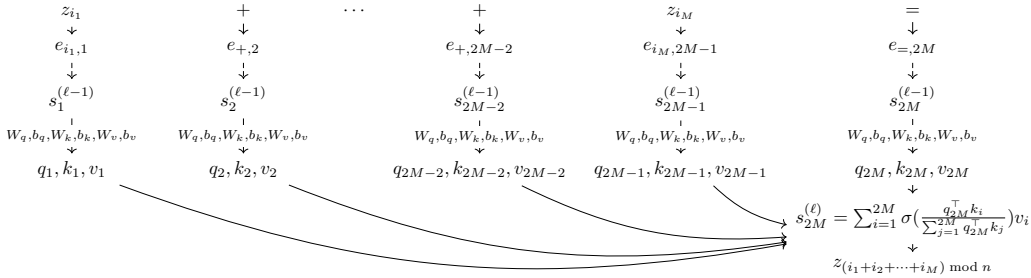


Figure 2: Illustration of the symbols defined for the hidden states of tokens and the variables for the attention layers

2.5 HOW DO LLMs LEARN ALGEBRAIC STRUCTURE?

In this section, we provide theoretical evidence that LLMs composed of attention layers can be constructed to compute addition under commutativity and identity. We present an example using \mathbb{Z}_n with M input elements as an illustrative example (see Fig. 2). The input sequence intersperses elements $z_i \in \mathbb{Z}_n$ with the symbols “+” and “=”, resulting in a total length of $2M$. We denote e_i , where $i \in \{0, \dots, n-1\}$ as the embedding for $z_i \in \mathbb{Z}_n$, and let e_+ and $e_=-$ represent the embeddings for the “+” and “=” tokens, respectively. If z_i appears in position m where $1 \leq m \leq 2M$, its embedding is $e_{i,m}$. Analogously, if a “+” or “=” symbol appears at position m , its embedding is $e_{+,m}$ or $e_{=,m}$. These embeddings consist of word embedding w_i and position embedding p_m , as $e_{i,m} = [w_i, p_m]^\top$ for $i \in \{0, 1, \dots, n-1\} \cup \{+, =\}$. We assume that all the vectors w_i and p_m are mutually orthogonal. In the following section, we illustrate how a language model can enforce commutativity for the “+” operator by providing a proof through construction. Specifically, we assign explicit values to the model’s weights and biases, assuming that these parameters can be learned from training. In the ℓ -th attention layer, we denote $s_m^{(\ell-1)}$ as the hidden state at position m from the previous attention layer. Let W_q, W_k, W_v and b_q, b_k, b_v denote the attention weights and biases that produce the query, key, and value, respectively, and let q_m, k_m , and v_m denote the query, key, and value vectors at position m in the ℓ -th attention layer, respectively. Furthermore, we define $s_{2M}^{(\ell)}$ as the hidden state in position $2M$ after ℓ attention layers. Namely, $s_{2M}^{(\ell)} = \sum_{i=1}^{2M} \sigma(\frac{q_{2M}^\top k_i}{\sum_{j=1}^{2M} q_{2M}^\top k_j}) v_i$. The following theorem demonstrates that LLM can learn hidden states to achieve commutativity.

Theorem 2.1 (Commutativity-Invariant to the Input Permutations). *Given the LLMs’ settings mentioned in Sec.2.5, there exists a special assignment of the weights and biases W_q, W_k, W_v and b_q, b_k, b_v and specific assignment of embeddings $e_{i,m}$, for $i \in \{0, 1, \dots, n-1\} \cup \{+, =\}$, such that $s_{2M}^{(\ell)}$ could be invariant to the permutation of input elements $z_{i_1}, z_{i_2}, \dots, z_{i_M} \in \mathbb{Z}_n$. However, this invariance holds only when the input contains commutative operators.*

Proof. The proof can be found in Sec. A.3.1. □

With the invariance of the hidden states $s_{2M}^{(\ell)}$, the subsequent layers of the transformer could serve as a classifier, mapping $s_{2M}^{(\ell)}$ to the token $z_{(i_1+i_2+\dots+i_M) \bmod n}$, and hence endow the addition operation with commutativity. In addition to commutativity, we present a theorem demonstrating how an LLM can produce hidden states that remain essentially unchanged under the insertion of identity elements.

Theorem 2.2 (Identity-Invariant to the Insertion of Identity Tokens). *Under the LLM settings in Sec. 2.5, let $s_{2M'}^{(\ell)}$, where $M' = M + 1$ denote the hidden state after inserting an identity token z_0 and an operator’s token into the input sequence. There exists a specific assignment of weights and biases W_q, W_k, W_v and b_q, b_k, b_v , together with particular embeddings $e_{i,m}$ for $i \in \{0, 1, \dots, n-1\} \cup \{+, =\}$, such that $s_{2M'}^{(\ell)}$ is equal to $s_{2M}^{(\ell)}$. However, this property is valid only when the input includes operators for which z_0 serves as the identity element.*

Proof. The proof can be found in Sec. A.3.2. □

With this theorem, a classifier can interpret $s_{2M'}^{(\ell)}$ as $s_{2M}^{(\ell)}$, which is already learned from base equation in the training set. This ensures that the appending z_0 does not alter the outcome, thus reflecting the identity property.

Remark 2.3 (Non-uniqueness of Weights and Bias Assignments). Note that the weights, biases, and embeddings described in the proof of these theorems represent only one possible configuration to achieve commutativity and identity; many others could also be valid. However, it is critical that these properties be triggered specifically by operators with the properties of commutativity and identity, rather than by the operand tokens themselves.

Here we discuss a solution in which commutativity and identity arise without the existence of operators’ embeddings.

Remark 2.4 (Trivial Solution of Embeddings). Language models may converge to a trivial embedding solution to achieve commutativity and identity. For instance, one might assign all non-identity tokens $\{z_1, z_2, \dots, z_{n-1}, +, =\}$ zero-valued position embeddings: $e_{i,m} = [w_i, 0]^\top$ for $i \in \{1, 2, \dots, n-1\}$.

$1\} \cup \{+, =\}$, and give the identity token z_0 zero-valued word and position embeddings: $e_{0,m} = [0, 0]^\top$. While this setup indeed satisfies both commutativity and identity, these properties are no longer tied to the operator itself. Consequently, the model fails to produce correct outputs for operators lacking commutativity and identity—such as \ominus , \triangleleft , and \triangleright —when using these same trivial embeddings.

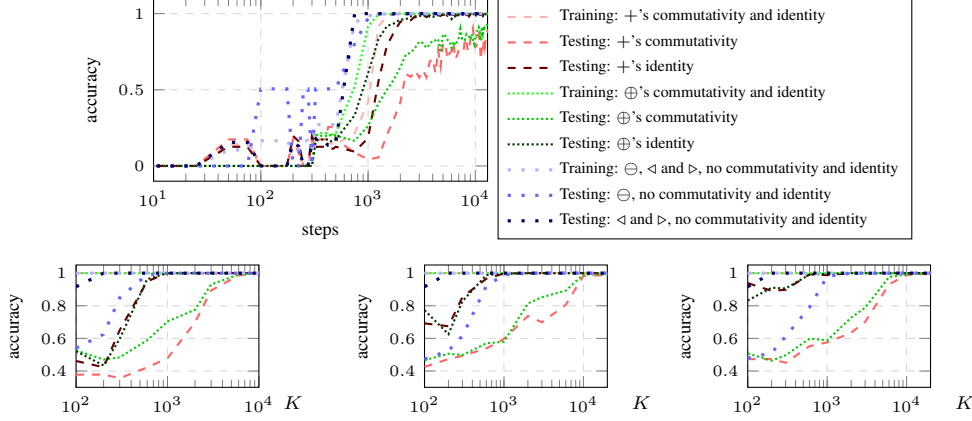


Figure 3: Plots of training and testing accuracy. The first row is the training dynamics for \mathbb{Z}_7 given the scale of training set $K = 3000$. The second row are the accuracies for \mathbb{Z}_7 (left), \mathbb{Z}_{11} (middle), \mathbb{Z}_{13} (right) with varying K of training set.

3 EXPERIMENTS

Settings: We conduct our experiments using the datasets described in Sec. 2.2, Sec. 2.3 and 2.4, which encompass addition problems that test for commutativity and identity of operator “+” and \oplus , as well as operations involving \ominus , \triangleleft , and \triangleright . We set $n = 7, 11$ and 13 for \mathbb{Z}_n and the number of input elements $M = 6$. For the language model, we choose GPT-2 Radford et al. (2019) but reinitialize its weights before training to strip away any pre-existing knowledge, ensuring that the model acquires its understanding of algebraic structures solely from our data. In addition, we customize the tokenizer so that each element is represented as a single token (e.g., z_{10} becomes the token `[z10]` rather than several character-based tokens). For reproducibility, we have made our experimental code publicly available¹.

Dataset Construction: We construct both training and testing sets by first choosing a scale K , which determines the number of examples. Each training set or testing set with scale K contains $10K$ instances, including

- $4K$ instances: Operators with commutativity and identity, including +’s commutativity, +’s identity, \oplus ’s commutativity, and \oplus ’s identity. Each of them contains K instances.
- $6K$ instances: Operator without commutativity and identity, including \ominus , \triangleleft , and \triangleright . Each of them contains $2K$ instances.

An example illustrating both training and testing with $K = 50$ appears in Sec. A.4. Throughout subsequent experiments, we fix the $K = 1000$ for the testing set, while K ranges from 100 to 30,000 for training set.

¹https://github.com/d09942015ntu/unraveling_llm_algebra

3.1 RESULTS

3.1.1 TRAINING DYNAMICS

We investigate how the training progresses for the case \mathbb{Z}_7 when $K = 3000$ for the training set. The upper row of Fig. 3 tracks the evolution of training and test accuracy over the course of training. We observe that the model ultimately achieves 100% accuracy in the training set, indicating that it has memorized all training instances. However, for the commutativity property of “+” or “ \oplus ” operators, it does not achieve high accuracy in testing set. A plausible explanation is that the scale of the training set is still insufficient. In the next experiment, we examine the testing accuracy for multiple training scales to investigate this further.

3.1.2 VARYING THE TRAINING SET’S SCALE.

We vary the size of the training set from $K = 100$ to $K = 30,000$ and measure testing accuracy once both the training and testing accuracy have plateaued. The results, depicted in the second row of Fig. 3, yield the following observations.

All Tasks Achieve Over 99% Testing Accuracy: We find that \triangleleft and \triangleright are the most easily learned, each achieving 99% testing accuracy with relatively few training samples. Next are \ominus and the identity properties of “+” and \oplus , which converge to 100% at around $K = 1000$. The most challenging part is learning the commutative properties of “+” and \oplus , which requires K between 10,000 and 20,000 to reach over 99% testing accuracy. Despite these differences, all tasks ultimately achieve over 99% accuracy, suggesting that commutativity and identity learning is indeed operator-driven rather than a trivial result of embeddings.

Generalization of Commutative Operations. Despite the number of training instances required to achieve high accuracy appears large, it is still much smaller than the full combinatorial space of expressions like $z_{i_1} + z_{i_2} + \dots + z_{i_6}$, which, for instance, includes $(7-1)^6 = 46,656$ possibilities in \mathbb{Z}_7 and $(13-1)^6 = 2,985,984$ in \mathbb{Z}_{13} . Thus, LLMs can actually learn and generalize commutativity for both “+” and \oplus without enumerating all possible permutations.

No Reliance on Numerical Computation We observe that “+” does not exceed \oplus in performance, indicating that LLMs learn commutativity and identity rather than relying on a direct numerical calculation. It also provides evidence that LLMs could not acquire computation skills for numerical values if the numerical values are not explicitly specified in the input.

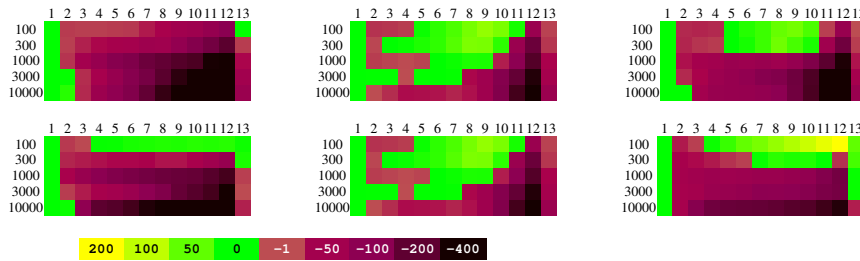


Figure 4: Visualization of S^{ℓ}_{com} and S^{ℓ}_{ide} where $1 \leq \ell \leq 13$. The upper row displays the values of $S^{\ell}_{\text{com}}(+, \ominus)$, $S^{\ell}_{\text{com}}(+, \triangleleft)$, and $S^{\ell}_{\text{com}}(+, \triangleright)$ and the lower row displays the values of $S^{\ell}_{\text{ide}}(+, \ominus)$, $S^{\ell}_{\text{ide}}(+, \triangleleft)$ and $S^{\ell}_{\text{ide}}(+, \triangleright)$. The numbers in the left axis represent $K \in \{100, 300, \dots, 10000\}$. For clarity, non-negative values are highlighted in green and yellow.

3.2 VISUALIZATION OF HIDDEN STATES

Commutative: As pointed out in Sec. 2.5, when the operator preserves commutativity, the hidden states remain invariant under permuting the inputs. In practice, however, these states need not be identical; it suffices that hidden states from different permutations be recognized as the same category. Consequently, given a set of input tokens, we expect slight variation (that is, a small

standard deviation) in the hidden states across different permutations. For example, assume that the inputs are $z_{i_1}, z_{i_2}, \dots, z_{i_m}$ and the output is y . Let \odot denote an operator. Then, considering all permutations, we have

$$y_1 = z_{i_1} \odot z_{i_2} \odot \dots \odot z_{i_m}, \quad y_2 = z_{i_2} \odot z_{i_1} \odot \dots \odot z_{i_3}, \quad \dots, \quad y_m = z_{i_m} \odot \dots \odot z_{i_2} \odot z_{i_1}.$$

If we define $s_i \in \mathbb{R}^D$ as the hidden state of each y_i where D is the size of hidden states, then for a commutative operator \odot , we would have $y_1 = y_2 = \dots = y_m$. Consequently, s_1, s_2, \dots, s_m should also be similar. The sum of their element-wise standard-deviation is defined by $S_{\text{std}}(s_1, \dots, s_m; \odot) = \sum_{k=1}^D \text{std}(\{s_1\}_k, \dots, \{s_m\}_k)$, where $\text{std}(x_1, \dots, x_m)$ denotes the standard deviation among x_1, \dots, x_m , $\{s_1\}_k$ denotes the k -th element of s_1 , and \odot indicates that the hidden states are produced by the operator \odot . For a non-commutative operator \odot' , these hidden states would differ more substantially, resulting in a higher value of $S_{\text{std}}(s_1, \dots, s_m; \odot')$. The left column of Fig. 4 shows the differences in S_{std} between the “+” operator and various non-commutative operators, denoted as:

$$S_{\text{com}}^{(\ell)}(+, \odot') = S_{\text{std}}(s_1^{(\ell)}, \dots, s_m^{(\ell)}; +) - S_{\text{std}}(s_1^{(\ell)}, \dots, s_m^{(\ell)}; \odot'), \text{ where } \odot' \in \{\ominus, \triangleleft, \triangleright\}.$$

Here, ℓ denotes the layer index (GPT-2 has 13 layers), and each column of the heat map in Fig. 4 corresponds to one of these layers. As the scale of the training set increases and the accuracy of the model in commutative operations improves, S_{std} for the commutative operator “+” becomes noticeably smaller compared to that of the non-commutative operators and consequently $S_{\text{com}}(+, \odot')$ become more negative.

Identity: We consider non-identity tokens z_1, z_2, \dots, z_m and an identity token z_0 to show the hidden states when the operator \odot remains invariant in the presence of an identity element. Concretely, we compare the outputs

$$\bar{y} = z_1 \odot z_2 \odot \dots \odot z_m, \quad y_1 = z_0 \odot z_1 \odot z_2 \odot \dots \odot z_m, \quad \dots, \quad y_m = z_1 \odot z_2 \odot \dots \odot z_m \odot z_0.$$

where \bar{y} is the result without z_0 . We denote \bar{s} and s_1, s_2, \dots, s_m as the hidden states of \bar{y} and y_1, y_2, \dots, y_m , respectively. If y_i for $i \in \{1, \dots, m\}$ and \bar{y} the same under the insertion of identity elements. Then, the distance between \bar{s} and any of s_i for $i \in \{1, \dots, m\}$ should be small. The sum of their distances is defined by $S_{\text{dist}}(\bar{s}, s_1, \dots, s_m; \odot) = \sum_{k=1}^D \sum_{i=1}^m |\{s_i\}_k - \{\bar{s}\}_k|$, where $\{s_i\}_k$ denotes the k -th element of s_i , and \odot indicates that the hidden states are produced by the operator \odot . For an operator without an identity element, denoted as \odot' , the distance between these two hidden states would be substantially larger, resulting in a larger value of $S_{\text{dist}}(\bar{s}, s_1, \dots, s_m; \odot')$. The right column of Fig. 4 shows the differences in S_{dist} between the “+” operator and various operators without identity elements, denoted as:

$$S_{\text{ide}}^{(\ell)}(+, \odot') = S_{\text{dist}}(\bar{s}^{(\ell)}, s_1^{(\ell)}, \dots, s_m^{(\ell)}; +) - S_{\text{dist}}(\bar{s}^{(\ell)}, s_1^{(\ell)}, \dots, s_m^{(\ell)}; \odot'), \text{ where } \odot' \in \{\ominus, \triangleleft, \triangleright\}.$$

As more training data is used and the model becomes better at identity-invariant operations, the value of S_{dist} for the “+” operator decreases significantly compared to operators without identity elements, leading to more negative values of $S_{\text{ide}}(+, \odot')$.

4 LIMITATIONS

In this work, we assume our problem scope is limited to a finite Abelian group \mathbb{Z}_5 , focusing exclusively on commutativity and identity. Other properties such as inverse and associativity, remain to be verified. Furthermore, real-world mathematical problems often involve real numbers and diverse forms of descriptions including natural language. Despite these limitations, we believe that our research takes the first step toward unraveling the mystery of the mathematical capabilities of LLMs. On the other hand, we tested only a relatively small LLM, GPT-2. Nevertheless, we hypothesize that larger models, with greater expressive power, are also capable of capturing algebraic structures within training data.

5 CONCLUSION

We have demonstrated that LLMs can learn and internalize fundamental algebraic properties, especially commutativity and identity, purely from training data. Our strategy involved constructing a

dataset of finite Abelian group expressions, ensuring that both commutative and identity instances appear in training and are held out for testing. Using a reinitialized GPT-2, we observed successful generalization to unseen tasks. Furthermore, We also provided a constructive proof showing how transformer-based models preserve invariance under permutations and identity insertion. Hidden-state visualizations revealed that operators preserving commutativity and identity produced more uniform internal representations compared to those that did not. Although our experiments centered on finite Abelian groups and basic algebraic properties, these results indicate the potential for LLMs to acquire and generalize more intricate algebraic structures directly from data. Extensions to larger systems, real numbers, advanced group properties, and more natural language settings remain promising directions for future research.

ACKNOWLEDGMENT

This work was supported in part by the Asian Office of Aerospace Research & Development (AOARD) under Grant NTU-112HT911020, National Science and Technology Council of Taiwan under Grant NSTC-112-2221-E-002-204- and NSTC-113-2221-E-002-208-, Ministry of Education (MOE) of Taiwan under Grant NTU-113L891406, and Ministry of Environment under Grant NTU-113BT911001

REFERENCES

- Anthropic. Claude 3 haiku: our fastest model yet. <https://www.anthropic.com/news/claude-3-haiku>, 2024. Accessed: 2024-10-21.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Chunyuan Deng, Zhiqi Li, Roy Xie, Ruidi Chang, and Hanjie Chen. Language models are symbolic learners in arithmetic. *arXiv preprint arXiv:2410.15580*, 2024.
- Zhifang Sui Fangwei Zhu, Damai Dai. Language models know the value of numbers, 2024.
- Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36, 2024.
- Antoine Gorceix, Bastien Le Chenadec, Ahmad Rammal, Nelson Vadori, and Manuela Veloso. Learning mathematical rules with large language models, 2024. URL <https://arxiv.org/abs/2410.16973>.
- Pei Guo, WangJie You, Juntao Li, Yan Bowen, and Min Zhang. Exploring reversal mathematical reasoning ability for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13671–13685, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36, 2024.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- Shima Imani and Hamid Palangi. Exploring group and symmetry principles in large language models. *arXiv preprint arXiv:2402.06120*, 2024.
- Pavan Karjol, Rohan Kashyap, and AP Prathosh. Neural discovery of permutation subgroups. In *International Conference on Artificial Intelligence and Statistics*, pp. 4668–4678. PMLR, 2023.

- Pavan Karjol, Rohan Kashyap, Aditya Gopalan, and AP Prathosh. A unified framework for discovering discrete symmetries. In *International Conference on Artificial Intelligence and Statistics*, pp. 793–801. PMLR, 2024.
- Junyu Lai, Jiahe Xu, Yao Yang, Yunpeng Huang, Chun Cao, and Jingwei Xu. Executing arithmetic: Fine-tuning large language models as turing machines. *arXiv preprint arXiv:2410.20124*, 2024.
- Amit Arnold Levy and Mor Geva. Language models encode numbers using digit representations in base 10. *arXiv preprint arXiv:2410.11781*, 2024.
- Tiedong Liu and Bryan Kian Hsiang Low. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. *arXiv preprint arXiv:2305.14201*, 2023.
- Yixin Liu, Avi Singh, C. Daniel Freeman, John D. Co-Reyes, and Peter J. Liu. Improving large language model fine-tuning for solving math problems. *arXiv preprint arXiv:2310.10047*, 2023.
- Zitao Liu, Ying Zheng, Zhibo Yin, Jiahao Chen, Tianqiao Liu, Mi Tian, and Weiqi Luo. Arithmeticgpt: Empowering small-size large language models with advanced arithmetic skills. *Machine Learning*, 114:24, 2025.
- Bohan Lyu, Yadi Cao, Duncan Watson-Parris, Leon Bergen, Taylor Berg-Kirkpatrick, and Rose Yu. Adapting while learning: Grounding llms for scientific problems with intelligent tool usage adaptation. *arXiv preprint arXiv:2411.00412*, 2024.
- Eran Malach. Auto-regressive next-token predictors are universal learners. *arXiv preprint arXiv:2309.06979*, 2023.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.
- OpenAI. Chatgpt. <https://chatgpt.com/>, 2024. Accessed: 2024-10-21.
- Ben Prystawski, Michael Li, and Noah Goodman. Why think step by step? reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, 36, 2024.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.
- Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. Numerologic: Number encoding for enhanced llms’ numerical reasoning. *arXiv preprint arXiv:2404.00459*, 2024.
- Si Shen, Peijun Shen, and Danhao Zhu. Revorder: A novel method for enhanced arithmetic in language models. *arXiv preprint arXiv:2402.03822*, 2024.
- Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis, 2023. URL <https://arxiv.org/abs/2305.15054>.
- Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. Mathscale: Scaling instruction tuning for mathematical reasoning. *arXiv preprint arXiv:2403.02884*, 2024.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Changnan Xiao and Bing Liu. A theory for length generalization in learning to reason. *arXiv preprint arXiv:2404.00560*, 2024.
- Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent space symmetry discovery. *arXiv preprint arXiv:2310.00105*, 2023.

David S. Yin and Xiaoxin Yin. Scaffolding learning: From specific to generic with large language models. *PLOS ONE*, 19(9):e0310409, 2024. URL <https://doi.org/10.1371/journal.pone.0310409>.

Wei Zhang, Wan Chaoqun, Yonggang Zhang, Yiu Ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. Interpreting and improving large language models in arithmetic calculation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 59932–59950, 2024.

Yongwei Zhou and Tiejun Zhao. Dual instruction tuning with large language models for mathematical reasoning. *arXiv preprint arXiv:2403.18295*, 2024.

A APPENDIX

A.1 RELATED WORKS

Theory of Chain-of-thought Reasoning in LLMs: Chain-of-Thought (CoT) techniques Wei et al. (2022) empower large language models (LLMs) to tackle complex mathematical reasoning tasks by breaking solutions into sequential steps, making them essential for solving mathematical problems. Recent studies shed light on CoT’s theoretical underpinnings. For example, Prystawski et al. (2024) models CoT with Bayesian networks, where questions, answers, and reasoning steps form interconnected nodes, demonstrating that structured reasoning improves LLM performance. Xiao & Liu (2024) introduces the concept of length generalization, showing that LLMs can extrapolate from simple examples to address more complex problems. Expanding the PAC learning framework, Malach (2023) shows that auto-regressive learners can effectively learn linear threshold circuits when CoT steps are provided. Additionally, Feng et al. (2024) proves that CoT enables transformers to handle dynamic programming problems, even with polynomially increasing complexity. Although these studies establish a theoretical basis for CoT, which decomposes complex mathematical problems into manageable steps, they rarely address how LLMs solve mathematical problems within a single step of CoT reasoning.

Enhancing mathematical reasoning in LLMs: Several recent works have developed different fine-tuning strategies to improve LLMs’ mathematical reasoning. First, Guo et al. (2024) mainly focuses on improving the “reversal curse” by introducing a reverse training task, thereby enhancing logical consistency. Similarly, Zhou & Zhao (2024) enhances the CoT ability by introducing two auxiliary tasks, including Intermediate Reasoning State Prediction and Instruction Reconstruction task, which model mathematical reasoning from both forward and reverse direction. Moreover, Liu et al. (2023) provides three fine-tuning methods to improve the LLMs’ performance on mathematical problems. By utilizing the supervision signal of the evaluation tasks, these methods effectively improve the model performance in generating solutions for math problems. Meanwhile, Yin & Yin (2024) proposes Scaffolding Learning, which first allows the model to master arithmetic operations and then fine-tunes it efficiently on the more general task of solving word math problems. Furthermore, Lyu et al. (2024) provides a two-component fine-tuning method, consisting of World Knowledge Distillation (WKD) and Tool Usage Adaptation (TUA). By leveraging these two components, the model surpasses state-of-the-art models such as GPT-4o in mathematical problem-solving. Finally, Tang et al. (2024) proposes a method called MathScale, which is used to construct the fine-tuning dataset MathScaleQA to enhance mathematical reasoning capabilities. Additionally, MWPBENCH is introduced as a benchmark to systematically evaluate performance. Although mathematical reasoning in these works has been enhanced, the underlying principles behind the reasoning process remain unknown. This gap suggests the need for further exploration in how LLMs solve such problems. To better understand these mechanism, we should start with the fundamental aspects, such as arithmetic, to uncover their underlying principles.

Improving Arithmetic ability in LLMs: The LLMs have demonstrated their power in natural language process tasks. However, they still exhibit limitations when it comes to performing arithmetic calculations. Recent studies have explored the application based on fine-tuning techniques to enhance the arithmetic capabilities of LLMs. For example, Liu et al. (2025) propose ArithmeticGPT, which enhances advanced arithmetic calculation, such as exponentiation, logarithms, and trigonometric functions. Similarly, Liu & Low (2023) propose supervised fine-tuning, mainly focuses on large-number arithmetic problem, particularly improving addition and developing decomposition strategies for multiplication and division. Nye et al. (2021) applied scratchpads fine-tuning, enabling the model to generalize to unseen 9-digit addition. In a different approach, Zhang et al. (2024) examines the inner component responsible for arithmetic calculations and uses the precise fine-tuning to enhance the attention head values and MLPs within the associated components. While Lai et al. (2024) fine-tunes LLMs to imitate Turing machine behavior, enabling step-by-step arithmetic calculations and enhancing their computational capability. Beyond fine-tuning, alternative methods have been proposed. Shen et al. (2024) apply RevOrder, a technique that reverses the arithmetic output order, to fine-tune LLMs, leading to a significant reduction in calculation errors. Schwartz et al. (2024) incorporates digit length information as a prefix, enabling the model to better understand numerical magnitude, thereby improving its arithmetic performance. Despite these improvements, these studies primarily aim to enhance arithmetic capabilities, rather than understanding the fundamental principle

of how LLMs could acquire arithmetic ability. Consequently, a deeper investigation into how LLMs internalize and generalize arithmetic concepts is still needed.

How Arithmetic Abilities Arise in LLMs: The mechanisms behind LLMs’ arithmetic abilities remain debated. Some studies suggest that LLMs encode numerical values internally. Fangwei Zhu (2024) demonstrates this by using linear probes on addition problems, showing that number values are encoded across layers and can be extracted. On the other hand, Levy & Geva (2024) finds that LLM errors are distributed across digits rather than numeric values, revealing that numbers are represented with per-digit circular structures in base 10. In addition, other works argue that LLMs rely on symbolic reasoning. Deng et al. (2024) shows LLMs learn simple patterns at the edges of a sequence of numbers faster than in the middle of the numbers of a sequence, indicating an easy-to-hard learning approach and symbolic arithmetic processing. Hanna et al. (2024) explores GPT-2 small’s mechanism for predicting valid end years in date-related tasks, identifying a circuit responsible for ”greater-than” comparisons that generalize across contexts. Additionally, Stolfo et al. (2023) demonstrates that LLMs transmit query-relevant information through attention mechanisms and process results with MLP modules, integrating them into the residual stream. Despite these insights, there is no consensus on whether LLMs primarily encode numerical values or rely on symbolic reasoning, highlighting the need for further research to clarify their mathematical processing mechanisms.

Machine Learning for Symmetric Discovery: The ability to discover symmetries enables machine learning models to uncover algebraic structures from training data. Karjol et al. (2023) demonstrate that sub-groups can be identified through a neural network with a specially designed architecture, supported by a general theorem. They validate their approach with numerical experiments on tasks such as image-digit sum and symmetric polynomial regression. Similarly, Karjol et al. (2024) present a unified framework for discovering symmetries across various subgroups, including locally symmetric, dihedral, and cyclic subgroups. Their architecture combines linear, matrix-valued, and non-linear functions to systematically capture invariance. Yang et al. (2023) introduce Latent LieGAN (LaLiGAN), a generative model that maps data to a latent space where nonlinear symmetries become linear. LaLiGAN simultaneously learns the mapping and the latent space symmetries, theoretically proving its ability to express nonlinear symmetries under specific group action conditions. However, these works do not explore the connection between symmetry learning and large language models’ (LLMs) arithmetic capabilities. While Imani & Palangi (2024) reveals that LLMs struggle with fundamental group properties and exhibit vulnerabilities in arithmetic reasoning, it does not investigate whether LLMs are possible to learn algebraic structures from training data. In contrast, our work demonstrates that LLMs can learn algebraic structures from training instances and generalize to solve unseen arithmetic problems.

LLMs can Learn Mathematical Rules: The work most closely related to ours is Gorceix et al. (2024), where the authors propose that LLMs can learn mathematical rules, such as distributivity or equation simplification. Although distributivity is also a type of algebraic structure, our work still has significant difference from them. First, our research demonstrates that LLMs can learn algebraic structures by training from scratch, showing that these rules are learned solely from the arithmetic equations we provide. We also rule out the possibility that LLMs learn these roles without any numerical computation. This differs from their approach, which relies on pre-trained models and cannot rule out the possibility that these rules were acquired from external materials or numerical computations. Additionally, our work further provides theoretical evidences of how transformers learn algebraic structures and providing an analysis of the hidden states of transformers, aspects that are not addressed in their study.

A.2 BACKGROUND KNOWLEDGE

Definition A.1. A finite Abelian group is a set \mathcal{G} with a binary operation \circ that satisfies the following properties:

- **Closure:** For any $a, b \in \mathcal{G}$, the results of $a \circ b$ are in \mathcal{G} .
- **Associativity:** For any $a, b, c \in \mathcal{G}$, $(a \circ b) \circ c = a \circ (b \circ c)$.
- **Commutativity:** For any $a, b \in \mathcal{G}$, $a \circ b = b \circ a$.

- **Identity:** There exists an element $e \in \mathcal{G}$ such that for any $a \in \mathcal{G}$, $a \circ e = e \circ a = a$.
- **Inverse:** For each $a \in \mathcal{G}$, there exists an element $a^{-1} \in \mathcal{G}$ such that $a \circ a^{-1} = a^{-1} \circ a = e$.

A.3 PROOF OF THEOREMS

A.3.1 PROOF OF THEOREM 2.1

Proof. Here, we prove this theorem by constructing a concrete example. In this example, the layer index $\ell = 1$, and hence the hidden state of the previous layer $s_m^{(\ell-1)}$ at position m is the input embedding $e_{i,m}$. We choose W_v to be the identity matrix, so the value vector at position m is simply:

$$v_m = W_v e_{i,m} = e_{i,m}.$$

To enforce commutativity for “+”, we assign the embedding of “+” at position m as

$$e_{+,m} = [w_+, -\infty]^\top,$$

where $-\infty$ represents the smallest floating-point value. Additionally, we set both W_q and W_k to zero matrices and make the biases b_q and b_k all ones. We assume that the context window size of attention is L , where $L \gg 2M$ and the remaining positions $2M < m \leq L$ are padded with zero embeddings

$$e_m = [0, 0]^\top \quad \text{for all } 2M < m \leq L.$$

Under these settings, the attention weights in the first layer become uniform across positions, i.e.,

$$\sigma\left(\frac{q_m^\top k_i}{\sum_{j=1}^L q_m^\top k_j}\right) = \frac{1}{L} \quad \text{for all } 1 \leq i, m \leq L.$$

Thus, the hidden state of the “=” token at position $2M$ after the first attention layer, denoted $s_{2M}^{(1)}$, is

$$s_{2M}^{(1)} = \sum_{i=1}^{2M} \sigma\left(\frac{q_{2M}^\top k_i}{\sum_{j=1}^{2M} q_{2M}^\top k_j}\right) v_i = \sum_{m=1}^{2M} \frac{1}{L} e_{i,m} = \left[\frac{1}{L} \sum_{m=1}^{2M} w_{i,m}, -\infty\right]^\top.$$

Since $\left[\frac{1}{L} \sum_{m=1}^{2M} w_{i,m}, -\infty\right]^\top$ is invariant to the position embeddings of e_{i_1}, \dots, e_{i_M} relative to e_+ , this hidden state does not depend on the order of the input tokens z_{i_1}, \dots, z_{i_M} . \square

A.3.2 PROOF OF THEOREM 2.2

Proof. Building on the setup from the previous section, let each non-identity token z_i ($i \neq 0$) at position m have the embedding

$$e_{i,m} = [w_i, 0, p_m]^\top \quad \text{for } i \in \{1, 2, \dots, n-1\},$$

where w_i and p_m are mutually orthogonal vectors. We then define the embedding of the identity token z_0 at position m as

$$e_{0,m} = [0, w_0, p_m]^\top,$$

and the embedding of the addition operator “+” at position m as

$$e_{+,m} = [0, -\infty, -\infty]^\top.$$

When we append z_0 with an extra “+” operator to the sequence of input tokens, the position of “=” become $2M' = 2M + 2$, and the hidden state after the attention layer at position $2M'$ is

$$s_{2M'}^{(1)} = \frac{\sum_{m=1}^{2M} e_{i,m} + e_{+,2M+1} + e_{0,2M+2}}{L} = \left[\frac{1}{L} \sum_{m=1}^{2M} w_{i,m}, -\infty, -\infty\right]^\top,$$

and the hidden states before inserting z_0 is

$$s_{2M}^{(1)} = \frac{\sum_{m=1}^{2M} e_{i,m}}{L} = \left[\frac{1}{L} \sum_{m=1}^{2M} w_{i,m}, -\infty, -\infty\right]^\top = s_{2M'}^{(1)}.$$

\square

A.4 EXAMPLE OF DATASET FOR \mathbb{Z}_7 WITH $K = 50$ Training, for Operator $+$'s Commutativity

$$\begin{array}{lll}
z_2 + z_2 + z_4 + z_3 + z_6 + z_4 = z_0 & z_4 + z_3 + z_6 + z_4 + z_2 + z_2 = z_0 & z_3 + z_6 + z_4 + z_4 + z_2 + z_2 = z_0 \\
z_6 + z_2 + z_3 + z_2 + z_4 + z_4 = z_0 & z_2 + z_3 + z_4 + z_2 + z_6 + z_4 = z_0 & z_2 + z_6 + z_4 + z_4 + z_3 + z_2 = z_0 \\
z_6 + z_2 + z_4 + z_2 + z_4 + z_3 = z_0 & z_4 + z_2 + z_6 + z_2 + z_3 + z_4 = z_0 & z_6 + z_2 + z_2 + z_3 + z_4 + z_4 = z_0 \\
z_6 + z_4 + z_2 + z_2 + z_3 + z_4 = z_0 & z_6 + z_6 + z_4 + z_4 + z_2 + z_3 = z_4 & z_6 + z_4 + z_2 + z_6 + z_4 + z_3 = z_4 \\
z_4 + z_6 + z_2 + z_3 + z_6 + z_4 = z_4 & z_4 + z_2 + z_6 + z_4 + z_6 + z_3 = z_4 & z_6 + z_4 + z_6 + z_3 + z_2 + z_4 = z_4 \\
z_4 + z_4 + z_6 + z_2 + z_3 + z_6 = z_4 & z_6 + z_3 + z_4 + z_6 + z_4 + z_2 = z_4 & z_3 + z_6 + z_2 + z_4 + z_4 + z_6 = z_4 \\
z_6 + z_3 + z_4 + z_4 + z_6 + z_2 = z_4 & z_6 + z_2 + z_4 + z_4 + z_3 + z_6 = z_4 & z_5 + z_5 + z_6 + z_2 + z_3 + z_3 = z_3 \\
z_5 + z_2 + z_3 + z_5 + z_6 + z_3 = z_3 & z_6 + z_2 + z_3 + z_5 + z_3 + z_5 = z_3 & z_5 + z_2 + z_3 + z_3 + z_5 + z_6 = z_3 \\
z_3 + z_5 + z_5 + z_2 + z_3 + z_6 = z_3 & z_5 + z_6 + z_3 + z_2 + z_3 + z_5 = z_3 & z_2 + z_5 + z_3 + z_3 + z_6 + z_5 = z_3 \\
z_6 + z_2 + z_3 + z_3 + z_5 + z_5 = z_3 & z_3 + z_5 + z_3 + z_6 + z_5 + z_2 = z_3 & z_6 + z_3 + z_2 + z_3 + z_5 + z_5 = z_3 \\
z_5 + z_5 + z_5 + z_3 + z_5 + z_2 = z_4 & z_5 + z_5 + z_2 + z_5 + z_3 + z_5 = z_4 & z_3 + z_5 + z_5 + z_2 + z_5 + z_5 = z_4 \\
z_5 + z_2 + z_5 + z_5 + z_3 + z_5 = z_4 & z_5 + z_3 + z_2 + z_5 + z_5 + z_5 = z_4 & z_3 + z_5 + z_5 + z_5 + z_2 + z_5 = z_4 \\
z_5 + z_5 + z_3 + z_5 + z_5 + z_2 = z_4 & z_5 + z_5 + z_5 + z_3 + z_2 + z_5 = z_4 & z_5 + z_2 + z_3 + z_5 + z_5 + z_5 = z_4 \\
z_2 + z_5 + z_3 + z_5 + z_5 + z_5 = z_4 & z_6 + z_6 + z_2 + z_5 + z_5 + z_3 = z_6 & z_6 + z_6 + z_5 + z_2 + z_5 + z_3 = z_6 \\
z_5 + z_6 + z_3 + z_2 + z_5 + z_6 = z_6 & z_5 + z_3 + z_2 + z_6 + z_6 + z_5 = z_6 & z_6 + z_5 + z_4 + z_1 + z_2 + z_4 = z_1 \\
z_6 + z_2 + z_3 + z_5 + z_6 + z_1 = z_2 & z_4 + z_6 + z_2 + z_2 + z_4 + z_5 = z_2 & z_2 + z_3 + z_6 + z_1 + z_5 + z_5 = z_1 \\
z_5 + z_4 + z_5 + z_5 + z_4 + z_2 = z_4 & z_4 + z_1 + z_1 + z_4 + z_5 + z_1 = z_2 &
\end{array}$$

Testing, for Operator $+$'s Commutativity

$$\begin{array}{lll}
z_4 + z_4 + z_5 + z_6 + z_2 + z_1 = z_1 & z_4 + z_2 + z_6 + z_5 + z_4 + z_1 = z_1 & z_4 + z_4 + z_2 + z_1 + z_6 + z_5 = z_1 \\
z_6 + z_2 + z_1 + z_5 + z_4 + z_4 = z_1 & z_6 + z_4 + z_1 + z_4 + z_2 + z_5 = z_1 & z_5 + z_2 + z_4 + z_4 + z_6 + z_1 = z_1 \\
z_4 + z_5 + z_4 + z_6 + z_1 + z_2 = z_1 & z_6 + z_5 + z_1 + z_4 + z_2 + z_4 = z_1 & z_1 + z_2 + z_4 + z_6 + z_5 + z_4 = z_1 \\
z_6 + z_1 + z_6 + z_2 + z_3 + z_5 = z_2 & z_2 + z_5 + z_3 + z_1 + z_6 + z_6 = z_2 & z_3 + z_2 + z_6 + z_1 + z_5 + z_6 = z_2 \\
z_6 + z_2 + z_5 + z_1 + z_3 + z_6 = z_2 & z_6 + z_6 + z_3 + z_5 + z_2 + z_1 = z_2 & z_1 + z_6 + z_5 + z_6 + z_3 + z_2 = z_2 \\
z_5 + z_1 + z_6 + z_3 + z_6 + z_2 = z_2 & z_5 + z_6 + z_6 + z_3 + z_1 + z_2 = z_2 & z_5 + z_3 + z_6 + z_2 + z_6 + z_1 = z_2 \\
z_2 + z_4 + z_5 + z_2 + z_4 + z_6 = z_2 & z_4 + z_5 + z_4 + z_2 + z_2 + z_6 = z_2 & z_2 + z_5 + z_2 + z_4 + z_4 + z_6 = z_2 \\
z_5 + z_2 + z_6 + z_4 + z_2 + z_4 = z_2 & z_5 + z_4 + z_6 + z_2 + z_4 + z_2 = z_2 & z_2 + z_4 + z_2 + z_6 + z_4 + z_5 = z_2 \\
z_4 + z_2 + z_4 + z_5 + z_6 + z_2 = z_2 & z_2 + z_4 + z_5 + z_6 + z_4 + z_2 = z_2 & z_4 + z_5 + z_2 + z_2 + z_4 + z_6 = z_2 \\
z_2 + z_1 + z_6 + z_5 + z_5 + z_3 = z_1 & z_6 + z_5 + z_3 + z_1 + z_5 + z_2 = z_1 & z_6 + z_2 + z_5 + z_5 + z_1 + z_3 = z_1 \\
z_3 + z_6 + z_1 + z_5 + z_2 + z_5 = z_1 & z_5 + z_2 + z_5 + z_1 + z_6 + z_3 = z_1 & z_3 + z_1 + z_6 + z_2 + z_5 + z_5 = z_1 \\
z_3 + z_6 + z_5 + z_2 + z_5 + z_1 = z_1 & z_6 + z_5 + z_1 + z_3 + z_2 + z_5 = z_1 & z_1 + z_5 + z_5 + z_2 + z_6 + z_3 = z_1 \\
z_2 + z_5 + z_5 + z_5 + z_4 + z_4 = z_4 & z_4 + z_5 + z_5 + z_2 + z_5 + z_4 = z_4 & z_2 + z_5 + z_5 + z_4 + z_5 + z_4 = z_4 \\
z_2 + z_5 + z_4 + z_5 + z_5 + z_4 = z_4 & z_5 + z_4 + z_5 + z_5 + z_2 + z_4 = z_4 & z_2 + z_4 + z_5 + z_4 + z_5 + z_5 = z_4 \\
z_5 + z_4 + z_2 + z_5 + z_5 + z_4 = z_4 & z_5 + z_4 + z_5 + z_2 + z_5 + z_4 = z_4 & z_5 + z_2 + z_4 + z_5 + z_5 + z_4 = z_4 \\
z_4 + z_1 + z_4 + z_5 + z_1 + z_1 = z_2 & z_1 + z_5 + z_4 + z_1 + z_1 + z_4 = z_2 & z_1 + z_1 + z_4 + z_1 + z_5 + z_4 = z_2 \\
z_4 + z_1 + z_1 + z_5 + z_4 + z_1 = z_2 & z_1 + z_5 + z_1 + z_4 + z_1 + z_4 = z_2 &
\end{array}$$

Training, for Operator $+$'s Identity

$$\begin{array}{lll}
z_0 + z_4 + z_3 + z_5 + z_3 + z_1 = z_2 & z_4 + z_0 + z_3 + z_5 + z_3 + z_1 = z_2 & z_4 + z_3 + z_0 + z_5 + z_3 + z_1 = z_2 \\
z_4 + z_3 + z_5 + z_0 + z_3 + z_1 = z_2 & z_4 + z_3 + z_5 + z_3 + z_0 + z_1 = z_2 & z_4 + z_3 + z_5 + z_3 + z_1 + z_0 = z_2 \\
z_4 + z_3 + z_5 + z_3 + z_1 = z_2 & z_0 + z_1 + z_5 + z_6 + z_6 + z_1 = z_5 & z_1 + z_0 + z_5 + z_6 + z_6 + z_1 = z_5 \\
z_1 + z_5 + z_0 + z_6 + z_6 + z_1 = z_5 & z_1 + z_5 + z_6 + z_0 + z_6 + z_1 = z_5 & z_1 + z_5 + z_6 + z_6 + z_0 + z_1 = z_5 \\
z_1 + z_5 + z_6 + z_6 + z_1 + z_0 = z_5 & z_1 + z_5 + z_6 + z_6 + z_1 = z_5 & z_0 + z_5 + z_2 + z_5 + z_2 + z_3 = z_3 \\
z_5 + z_0 + z_2 + z_5 + z_2 + z_3 = z_3 & z_5 + z_2 + z_0 + z_5 + z_2 + z_3 = z_3 & z_5 + z_2 + z_5 + z_0 + z_2 + z_3 = z_3 \\
z_5 + z_2 + z_5 + z_2 + z_0 + z_3 = z_3 & z_5 + z_2 + z_5 + z_2 + z_3 + z_0 = z_3 & z_5 + z_2 + z_5 + z_2 + z_3 = z_3 \\
z_0 + z_3 + z_1 + z_2 + z_2 + z_2 = z_3 & z_3 + z_0 + z_1 + z_2 + z_2 + z_2 = z_3 & z_3 + z_1 + z_0 + z_2 + z_2 + z_2 = z_3 \\
z_3 + z_1 + z_2 + z_0 + z_2 + z_2 = z_3 & z_3 + z_1 + z_2 + z_2 + z_0 + z_2 = z_3 & z_3 + z_1 + z_2 + z_2 + z_2 + z_0 = z_3 \\
z_3 + z_1 + z_2 + z_2 + z_2 = z_3 & z_0 + z_2 + z_4 + z_4 + z_3 + z_4 = z_3 & z_2 + z_0 + z_4 + z_4 + z_3 + z_4 = z_3 \\
z_2 + z_4 + z_0 + z_4 + z_3 + z_4 = z_3 & z_2 + z_4 + z_4 + z_0 + z_3 + z_4 = z_3 & z_2 + z_4 + z_4 + z_3 + z_0 + z_4 = z_3 \\
z_2 + z_4 + z_4 + z_3 + z_4 + z_0 = z_3 & z_2 + z_4 + z_4 + z_3 + z_4 = z_3 & z_0 + z_1 + z_4 + z_6 + z_5 + z_2 = z_4 \\
z_1 + z_0 + z_4 + z_6 + z_5 + z_2 = z_4 & z_1 + z_4 + z_0 + z_6 + z_5 + z_2 = z_4 & z_1 + z_4 + z_6 + z_0 + z_5 + z_2 = z_4 \\
z_1 + z_4 + z_6 + z_5 + z_0 + z_2 = z_4 & z_1 + z_4 + z_6 + z_5 + z_2 = z_4 & z_5 + z_6 + z_1 + z_4 + z_4 = z_6 \\
z_4 + z_2 + z_4 + z_6 + z_3 = z_5 & z_5 + z_1 + z_1 + z_5 + z_3 = z_1 & z_2 + z_1 + z_2 + z_6 + z_2 = z_6 \\
z_6 + z_1 + z_2 + z_5 + z_4 = z_4 & z_1 + z_4 + z_6 + z_1 + z_3 = z_1 & z_4 + z_1 + z_2 + z_4 + z_6 = z_3 \\
z_4 + z_4 + z_1 + z_2 + z_2 = z_6 & z_2 + z_1 + z_3 + z_5 + z_4 = z_1 &
\end{array}$$

Testing, for Operator $+$'s Identity

$$\begin{array}{lll}
z_0 + z_5 + z_6 + z_1 + z_4 + z_4 = z_6 & z_5 + z_0 + z_6 + z_1 + z_4 + z_4 = z_6 & z_5 + z_6 + z_0 + z_1 + z_4 + z_4 = z_6 \\
z_5 + z_6 + z_1 + z_0 + z_4 + z_4 = z_6 & z_5 + z_6 + z_1 + z_4 + z_0 + z_4 = z_6 & z_5 + z_6 + z_1 + z_4 + z_4 + z_0 = z_6
\end{array}$$

$$\begin{array}{llll}
z_2 \triangleleft z_4 \triangleleft z_4 \triangleleft z_0 \triangleleft z_3 \triangleleft z_4 = z_2 & z_2 \triangleleft z_4 \triangleleft z_4 \triangleleft z_3 \triangleleft z_0 \triangleleft z_4 = z_2 & z_2 \triangleleft z_4 \triangleleft z_4 \triangleleft z_3 \triangleleft z_4 \triangleleft z_0 = z_2 & \\
z_2 \triangleleft z_4 \triangleleft z_4 \triangleleft z_3 \triangleleft z_4 = z_2 & z_0 \triangleleft z_1 \triangleleft z_4 \triangleleft z_6 \triangleleft z_5 \triangleleft z_2 = z_0 & z_1 \triangleleft z_0 \triangleleft z_4 \triangleleft z_6 \triangleleft z_5 \triangleleft z_2 = z_1 & \\
z_1 \triangleleft z_4 \triangleleft z_0 \triangleleft z_6 \triangleleft z_5 \triangleleft z_2 = z_1 & z_1 \triangleleft z_4 \triangleleft z_6 \triangleleft z_0 \triangleleft z_5 \triangleleft z_2 = z_1 & z_1 \triangleleft z_4 \triangleleft z_6 \triangleleft z_5 \triangleleft z_0 \triangleleft z_2 = z_1 & \\
z_2 \triangleleft z_4 \triangleleft z_6 \triangleleft z_5 \triangleleft z_2 = z_1 & z_5 \triangleleft z_6 \triangleleft z_1 \triangleleft z_4 \triangleleft z_4 = z_5 & z_4 \triangleleft z_2 \triangleleft z_4 \triangleleft z_6 \triangleleft z_3 = z_4 & z_5 \triangleleft z_1 \triangleleft z_1 \triangleleft z_5 \triangleleft z_3 = z_5 \\
z_2 \triangleleft z_1 \triangleleft z_2 \triangleleft z_6 \triangleleft z_2 = z_2 & z_6 \triangleleft z_1 \triangleleft z_2 \triangleleft z_5 \triangleleft z_4 = z_6 & z_1 \triangleleft z_4 \triangleleft z_6 \triangleleft z_1 \triangleleft z_3 = z_1 & z_4 \triangleleft z_1 \triangleleft z_2 \triangleleft z_4 \triangleleft z_6 = z_4 \\
z_4 \triangleleft z_4 \triangleleft z_1 \triangleleft z_2 \triangleleft z_2 = z_4 & z_2 \triangleleft z_1 \triangleleft z_3 \triangleleft z_5 \triangleleft z_4 = z_2 & &
\end{array}$$

Testing, for Operator \triangleleft

[illegible]

Training, for Operator ▷

[illegible]

[illegible]

Testing, for Operator \triangleright

[illegible]