An Investigation of Neuron Activation as a Unified Lens to Explain Chain-of-Thought Eliciting Arithmetic Reasoning of LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown strong arithmetic reasoning capabilities when prompted with Chain-of-Thought (CoT) prompts. However, we have only a limited understanding of how they are processed by LLMs. To demystify it, prior work has primarily focused on ablating different components in the CoT prompt and empirically observing their resulting LLM performance change (Madaan and Yazdanbakhsh, 2022; Wang et al., 2023; Ye et al., 2023). Yet, the reason why these components are important to LLM reasoning is not explored. To fill this gap, in this work, we investigate "neuron activation" as a lens to provide a unified explanation to observations made by prior work. Specifically, we look into neurons within the feed-forward layers of LLMs that may have activated their arithmetic reasoning capabilities, using Llama2 (Touvron et al., 2023) as an example. To facilitate this investigation, we also propose an approach based on GPT-4 to automatically identify neurons that imply arithmetic reasoning. Our analyses revealed that the activation of reasoning neurons in the feed-forward layers of an LLM can explain the importance of various components in a CoT prompt, and future research can extend it for a more complete understanding.[1]

1 Introduction

Arithmetic reasoning is one of the emergent properties in large language models (LLMs), which is necessary for them to tackle tasks that require multiple steps to arrive at the correct answer. In recent years, Chain-of-Thought (CoT) has become a popular prompting strategy to elicit reasoning[2] in LLMs (Wei et al., 2022). Despite its successes, there is little understanding of what makes it effective and how LLMs utilize it to facilitate reasoning.

To address this concern, a line of research has focused on decomposing the CoT prompt into various components and performing ablation studies on them to ascertain the significance of each component on the LLM reasoning performance (Madaan and Yazdanbakhsh, 2022; Wang et al., 2023; Ye et al., 2023). Although these studies have yielded several insightful observations on the effect of input on LLM's reasoning performance, they do not shed light on how these inputs are being processed internally by LLMs to perform reasoning.

On the other hand, there is a growing body of research in the field of mechanistic interpretability (Elhage et al., 2021; Wang et al., 2022a) that specifically examines the internals of LLMs to understand their mechanism. In this vein, Stolfo et al. (2023) studied the internal mechanism of LLMs to perform arithmetic calculation, suggesting that attention heads facilitate information traversal, while the feed-forward layer (FFN) handles information processing to produce accurate answers for a given computation. However, Stolfo et al. (2023) only studied the mechanism for a single mathematical computation and doesn't study arithmetic reasoning in full scope. In parallel, some other research demonstrated that LLMs consist of neurons that can be associated with human-interpretable concepts, which play a crucial role in various capabilities of LLMs (Geva et al., 2022; Dai et al., 2021; Gurnee et al., 2024). Specifically, Geva et al. (2022) showed that neurons in the FFN layer of a transformer model (Vaswani et al., 2017) form key-value pairs that facilitate next-token prediction by promoting concepts in the vocabulary space. However, none of the prior work has applied the intuition to understand LLM reasoning.

Motivating by the need to form a deeper understanding of how CoT prompts elicit reasoning in LLMs and observing the pivotal role of neurons within the FFN layers of LLMs, in this work, we propose to investigate the activation of FFN neu-

---

[1] Our source code will be released upon paper acceptance.
[2] Our work focuses on "arithmetic reasoning". For ease of presentation, we use "reasoning" interchangeably with it.

rons in LLMs as a lens to interpret their arithmetic reasoning capabilities. In particular, we aim to use neuron activation to provide a unified explanation of observations that were only empirically made by prior work (Madaan and Yazdanbakhsh, 2022; Wang et al., 2023; Ye et al., 2023).

To this end, we first propose an approach based on GPT-4 (OpenAI, 2023) to *automatically* search for neurons that are related to arithmetic reasoning (e.g., arithmetic addition, logical connections, etc.). Prior work trying to search for concept-relevant neurons has fully relied on human labor. For example, Geva et al. (2022) introduced an approach to manually examine a neuron's top promoted tokens and determine if the neuron promotes the given concept or not. However, this manual approach can become impractical for LLMs with a large number of layers and numerous neurons per layer. Our approach instead decides whether a given neuron expresses a certain concept automatically by prompting the GPT-4 to read its top promoted tokens and make a judgment on its represented concept. Our experimental results demonstrate the high effectiveness of utilizing GPT-4 for this purpose. Subsequently, we apply our proposed approach to identify FFN neurons in Llama2-7B that promote concepts relevant to arithmetic reasoning.

Leveraging the identified reasoning neurons, we performed a series of analyses on observations made by prior work (Madaan and Yazdanbakhsh, 2022; Wang et al., 2023; Ye et al., 2023), including the importance of textual explanation, equations, arithmetic diversity, and the negligible impact of incorrect labels in CoT prompts. Specifically, we analyzed the activation patterns of the identified reasoning neurons such as their activation frequency and strength to gain insights into prior observations. Our results reveal that the activation of FFN neurons in LLMs can be used to explain their arithmetic reasoning capability. We then conclude the paper with a discussion of future work that can complement the proposed neuron activation analysis with other approaches to form a more complete understanding of LLM reasoning.

2 Background and Related Work

2.1 Prior Work towards Understanding the CoT Reasoning of LLMs

Prior studies attempted to understand the arithmetic reasoning in LLMs by decomposing the Chain-of-Thought (CoT) prompt into different semantic components and evaluating their importance via ablation studies. We present a summary of the major findings from prior work in Table 1. For example, to understand whether equations matter in the few-shot CoT prompt, Ye et al. (2023) experimented with a CoT variant where all equations (e.g., "21 - 15 = 6") were eliminated and only the calculation results (e.g., "6") was presented. By observing the resulting LLM performance change, one can empirically gauge the importance of equations in a CoT prompt. While previous studies have highlighted the significance of various components (e.g., textual explanation, equations, etc.) within the CoT prompt, the underlying reason behind these observations remains unanswered. This thus motivates us to provide a more fundamental understanding of the inner mechanism of LLM reasoning.

2.2 Interpreting Neurons of LLMs

Many prior interpretability works have studied neurons to understand the inner mechanism of LLMs and have led to the discovery of many interesting types of neurons such as knowledge neurons (Dai et al., 2021), skill neurons (Wang et al., 2022b), sentiment neurons (Radford et al., 2017), concept neurons (Geva et al., 2022), universal neurons (Gurnee et al., 2024), and many others related to linguistic and grammar features (Durrani et al., 2022; Sajjad et al., 2022). Furthermore, the activation patterns of these neurons have been found to significantly influence the behavior of LLMs (Geva et al., 2022). To discover the targeted neurons, probing is the most widely used approach, which involves training a simple classifier (probe) on the representations of neurons using a human-annotated dataset (Gurnee et al., 2023; Belinkov, 2022). Another popular approach specific to transformer-based LLMs is the projection of neuron representations to the vocabulary space, introduced by Geva et al. (2022), and has been widely adopted (Dar et al., 2022; Belrose et al., 2023; Ghandeharioun et al., 2024). However, to the best of our knowledge, none of the prior work has applied neuron activation to understand LLM reasoning. Our work draws inspiration from Geva et al. (2022) but extends it for a unified explanation of observations in CoT prompting. To this end, we also proposed an automatic approach based on GPT-4 for neuron discovery.

Relevant to our work, Stolfo et al. (2023) have also attempted to understand arithmetic reasoning by interpreting their neuron behaviors. However,

Research Questions	Examples in CoT Prompts	Prior Work	Findings
Does equation matter? (RQ3)	**w Equation:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been **21 - 15 = 6** trees. The answer is 6. **w/o Equation:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been **6** trees. The answer is 6.	Wang et al. (2023); Ye et al. (2023); Madaan and Yazdanbakhsh (2022)	Yes
Does textual explanation matter? (RQ4)	**w Textual Explanation:** Let's think step by step. First Leah had 32 chocolates and her sister had 42 chocolates. So in total they had 32 + 42 = 74 chocolates. Then they ate 35 chocolates. So there must be 74 - 35 = 39 chocolates. The answer is 39. **w/o Textual Explanation:** 32 + 42 = 74. 74 - 35 = 39. The answer is 39.	Wang et al. (2023); Ye et al. (2023); Madaan and Yazdanbakhsh (2022)	Yes
Does the diversity of arithmetic operators matter? (RQ5)	**AddOnly:** Let's think step by step. First there are 3 cars. Then 2 more cars arrive. So there must be 3 + 2 = 5 cars. The answer is 5. **MultOnly:** Let's think step by step. First a farmer has 5 cows. Then each cow has 4 legs. So the cows have 5 x 4 = 20 legs in total. The answer is 20.	Ye et al. (2023)	Yes
Does incorrect reasoning or gold label not matter? (RQ6)	**Correct Label:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6 trees. The answer is 6. **Incorrect Label:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = **1** trees. The answer is **1**. **OOD Label:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = **Dawson** trees. The answer is **Dawson**.	Wang et al. (2023); Ye et al. (2023)	No

Table 1: Summary of shared findings from prior works. Our reproduced results are shown in Table 2.

the majority of their study focused on coarser units such as the entire attention or FFN block. Furthermore, their investigation solely focused on how LLMs execute arithmetic calculations whereas the (multi-step) reasoning process is underexplored.

2.3 Concept Promotion via Neuron Activation of Geva et al. (2022)

Our work builds upon the findings of Geva et al. (2022), which shows the role of the feed-forward network (FFN) layer in the construction of an LLM's prediction – (a) each FFN layer induces *additive updates* to token representations, which can be further decomposed into weighted collections of *sub-updates*; (b) both the token representation and sub-updates of the FFN layer can be projected at any stage to a distribution over the output vocabulary. Through the vocabulary space projection, the authors found that the sub-updates of an FFN layer often encode human-interpretable concepts. Next, we briefly describe Geva et al. (2022)'s findings; more details should refer to the original paper.

Consider an auto-regressive transformer-based LLM consisting of L layers, which predicts the next token by projecting its last-layer hidden state onto a vocabulary $\mathcal{V}$ via an embedding matrix $E \in \mathbb{R}^{d \times |\mathcal{V}|}$, where d denotes the embedding size and $|\mathcal{V}|$ represents the vocabulary size. We denote the FFN component in the l-th layer as FFN^l. Given a token sequence $X = (x_1, ..., x_{|X|})$ as input, the representation of each token x_i at layer l (denoted as $x_i^l \in \mathbb{R}^d$) is updated by FFN^l as follows:

$$\bar{x}_i^l = x_i^l + FFN^l(x_i^l) \tag{1}$$

The updated representation $\bar{x}_i^l$ then goes through the multi-head self-attention at layer l, which results in x_i^{l+1} for the next FFN layer (i.e., FFN^{l+1}). With the residual connection (He et al., 2016), each FFN update can be seen as producing *additive updates* to the token representation.

In transformers, each FFN^l is defined with two parameter matrices $K^l, V^l \in \mathbb{R}^{d_m \times d}$, where d_m is the intermediate hidden dimension, and a non-linearity function f:

$$FFN^l(x_i^l) = f(K^l x_i^l)V^l \tag{2}$$

Eqn 2 can further be decomposed as:

$$FFN^l(x_i^l) = \sum_{j=1}^{d_m} f(x_i^l \cdot k_j^l)v_j^l = \sum_{j=1}^{d_m} m_{ij}^l v_j^l \tag{3}$$

where $k_j^l \in \mathbb{R}^d$ and $v_j^l \in \mathbb{R}^d$ are the j-th row of K^l and V^l, respectively, and $m_{ij}^l = f(x_i^l \cdot k_j^l)$ is a scalar representing the *activation coefficient* of v_j^l (i.e., the *neuron*). Geva et al. (2022) interpreted each term in this sum as a set of d_m *sub-updates* to the token representation. They also proposed to project this sub-update to the vocabulary by Ev_j^l. By analyzing the projected vocabulary tokens (typically tokens with top projection scores), they found that the sub-update often encodes human-interpretable concepts. It is important

to note that every v_j^l is a static parameter that is input-independent, while the coefficient m_{ij}^l depends on the input token x_i.

Observing their critical roles and leveraging their interpretability after projection, Geva et al. (2022) demonstrated the potential of encouraging non-toxic language by manipulating the coefficients of FFN neurons in LLMs. This was achieved by identifying FFN neurons representing non-toxic language concepts and then increasing their coefficients. Getting inspired by their findings, our work aims to explore: Can FFN neuron activation be similarly used to interpret and even control LLM reasoning? It is important to note that "toxicity" and "reasoning" represent distinct extents of abstraction. While whether a sentence is toxic or not can be judged by superficial keyword searching, "reasoning" is more abstract and can encompass multiple aspects (e.g., logical induction, mathematical calculation, etc.), which thus presents a significant challenge.

3 Neuron Discovery using GPT-4

To facilitate the neuron analysis, we first propose an approach for discovering neurons that express concepts related to arithmetic reasoning. To achieve the same goal, Geva et al. (2022) manually examined the top-scoring vocabulary tokens projected by each neuron v_j^l and annotated its concept. However, this manual search approach can become impractical for LLMs with deep layers and numerous sub-updates per layer. To overcome this inefficiency, we propose a method that leverages GPT-4 to automate the search process.

Our proposed approach involves two steps. First, for a given LLM, we store the K neurons v_j^l's with the largest coefficient m_{ij}^l from each layer l and at each generation time step i, using a set of examples $\mathcal{E}$ that showcase the LLM's capability (i.e., arithmetic reasoning in our case) to provide the prompt. We only considered the top-K neurons to narrow our search to the most activated neurons. This returns a set of candidate neurons $\mathcal{N}$. We present this step in Algorithm 1.

In the second step, we task GPT-4 to determine whether each neuron in $\mathcal{N}$ promotes a predefined concept C_{name} (e.g., arithmetic addition). However, employing GPT-4 to classify all neurons in $\mathcal{N}$ still requires a large number of prompts and may incur significant costs. To address this issue, we propose to first filter out the irrelevant neurons

---

Algorithm 1 Candidate Neuron Collection

1: **Input:** A set of examples $\mathcal{E}$ implying the capability, a filtering threshold K, the target LLM
2: **Output:** A set of candidate neurons $\mathcal{N}$.
3: Initialize $\mathcal{N} \leftarrow \{\}$
4: **for** each example in $\mathcal{E}$:
5:      **for** each decoding step i :
6:          **for** each layer $l = 1, ..., L$:
7:              $\{m_{ij'}^l\}_{j'=1}^K \leftarrow$ FindLargestK($\{m_{ij}^l\}_{j=1}^{d_m}, K$)
8:              $\mathcal{N} \leftarrow \mathcal{N} \cup \{v_j^l | m_{ij}^l \in \{m_{ij'}^l\}_{j'=1}^K\}$

Algorithm 2 Neuron Annotation via GPT-4

1: **Input:** Concept C_{name}, a set of seed tokens $\mathcal{S}_{name}$, filtering thresholds P and F, embedding E of LLM, and candidate neuron set $\mathcal{N}$.
2: **Output:** A subset of neurons $\mathcal{R} \subset \mathcal{N}$ representing concept C_{name}.
3: Initialize $\mathcal{R} \leftarrow \{\}$
4: **for** each neuron $v_n \in \mathcal{N}$:
5:      $\mathcal{V}_P = \{w_1, ..., w_P\} \leftarrow$ GetLargestP(Ev_n, P)
6:      **if** $|\mathcal{V}_P \cap \mathcal{S}_{name}| \geq F$:
7:          **if** GPT4ConceptQuery($\mathcal{V}_P, C_{name}$) :
8:              $\mathcal{R} \leftarrow \mathcal{R} \cup \{v_n\}$

---

by using a set of human-annotated "seed tokens" (denoted as $\mathcal{S}_{name}$) that are likely to be associated with the given concept as per human intuition. For instance, when searching for neurons that promote arithmetic addition, relevant tokens may include "add", "addition", "sum", "+", and "plus". Although a neuron that promotes the given concept may not invariably promote all the tokens from the seed tokens, it is quite probable that it promotes at least some of them. Leveraging this insight, we filter out neurons that do not consist of at least a threshold of F seed tokens in their top-P promoted tokens $\mathcal{V}_P$, obtained by projecting the neuron to vocabulary space. Finally, we prompt GPT-4 to inquire whether a neuron from filtered N promotes a given concept or not, This step is described in Algorithm 2, and we include the prompt script in Appendix A.

Neuron Activation Following Geva et al. (2022), we consider a neuron being *activated* in a layer l at a time step i when the neuron's coefficient m_{ij}^l is ranked at top 10. The other alternative would be to define a threshold based on m_{ij}^l to determine its activation. However, coming up with an appropriate threshold poses a challenge, as the threshold value may vary across different layers or even among the individual neurons. Consequently, we opt to focus solely on neurons with the top 10 largest coefficients in our analysis.

CoT Prompt	Accuracy
CoT	16.83%
w/o Equation (RQ3)	12.58%
w/o Textual Explanation (RQ4)	13.41%
AddOnly (RQ5)	13.26%
MultOnly (RQ5)	13.13%
Incorrect Label (RQ6)	16.45%
OOD Label (RQ6)	7.58%

Table 2: The accuracy of Llama2-7B on GSM8k test set based on different CoT prompts.

4 Experimental Setup

Dataset and Model Setup We conduct our experiment on the GSM8k dataset (Cobbe et al., 2021), which is widely used for evaluating the arithmetic reasoning capabilities of LLMs. It consists of diverse grade school math word problems and only requires basic arithmetic operations to solve, often involving problem-solving steps ranging from two to eight. We use Llama2-7B (Touvron et al., 2023) as our model to investigate the reasoning capabilities in LLMs. However, we believe that our findings apply to other transformer-based decoder-only LLMs as well.

We base our experiments on the CoT prompts obtained from Fu et al. (2023), with a slight modification to encourage consistent format in multi-step reasoning, for the ease of further analysis. Each CoT prompt consists of eight exemplars. Additionally, we adapt the CoT prompts into different various to reproduce prior observations. For reproducibility purposes, we provide a complete list of our prompts in the Appendix H.

Before investigating the mechanism of LLM reasoning, we have conducted experiments to replicate and validate observations made by prior work (Table 1). For RQ4 and RQ6, different prior work adopted different ablation designs. We opted for the most suitable and fair design among them. The experimental results based on Llama2-7B are presented in Table 2, which present consistent observations as prior research. We refer readers to Appendix B for more details.

Summary of Research Questions (RQs) Leveraging the lens of neuron activation, we aim to answer two sets of questions. The first set of questions (RQs 1-2) tries to understand the underlying mechanism of LLM reasoning where we initially find different neurons related to arithmetic reasoning and explore the importance of these discovered reasoning neurons for activating reasoning in LLMs. Built upon this foundational understanding of LLMs' reasoning mechanism, the second set of questions (RQs 3-6) attempts to provide a unified explanation of observations made by prior work.

5 Understanding the Mechanism of Reasoning in LLMs

5.1 RQ1: Are there neurons or sub-updates related to the concept of "reasoning"?

To answer this question, we apply the proposed approach in Section 3 to automatically identify neurons implying a set of 7 concepts, including logical connectors, which plays a crucial role in deciding the reasoning direction, a set of four arithmetic operations (i.e., add, subtract, multiply, and division), and others (equals to and calculation), which are also important to arithmetic reasoning. Though they may not fully encompass arithmetic reasoning, these concepts are sufficient for an initial investigation of neuron activation. The seed token set S_{name} for each concept, the identified neuron examples, and the expanded concept tokens found in the identified neurons, are presented in Table 3. The specific implementation details are included in Appendix C.

We find a total of 113 neurons associated with the listed concept in Llama2-7B. We performed manual validation of the results and didn't find any objection. Notably, we discovered neurons that group certain concepts using different language characters. For instance, the neuron, "L21N7027", corresponding to the 21st layer and 7027-th row of V^{21}, promotes tokens like "and" and "+" with their corresponding translation for Chinese (U+4e0e) and Japanese (U+3068). Additionally, we also found some neurons with polysemantic characteristics, where a single neuron promotes multiple concepts. For instance, "L27N10751" promotes tokens related to both addition (+, plus, +=, ..) and subtraction (-, minus, -+, ..).

Activation Pattern of Reasoning Neurons Our further investigation found out intriguing activation pattern of reasoning neurons throughout an LLM's reasoning process. For example, in Figure 2 of Appendix D, we showed that the logical connector neurons are often activated at the beginning of a generated sentence, whereas arithmetic neurons are mostly activated in response to arithmetic symbols and numbers. Once a neuron is activated, it remains activated for a few subsequent time steps. This persistence implies a lasting impact of activated neurons on text generation in its proximity.

Concept	Seed Tokens	Expanded Tokens	#of Neurons	Exemplar Neurons
Logical Connectors (C_{logic})	{first, so, meaning, therefore, then, next, hence }	{logic, implies, thus, however, accordingly, subsequently, later, corresponding, etc. }	65	L10N9818{then, THEN, Then, then, ..}, L11N3000{therefore, Therefore, accordingly, donc,..}, L11N7742, L12N1030
Addition (C_{add})	{add, addition, +, sum, plus }	{added, U+002B, adding, ++, increment, total, etc.}	18	L12N4814{added ,addition ,add,..}, L21N7027{+ ,add ,U+4e0e, U+306,..}, L27N10751{+, plus, -, minus,..}
Subtraction (C_{sub})	{subtract, -, minus, sub }	{ -=, negative, U+2212, etc. }	2	L19N7900{-= ,- ,minus,2212, ..}, L25N5227
Multiplication (C_{mul})	{multiply, product, times, mult, *, x }	{ multip, multi, U+00D7, double, twice, triple, fold, larger, etc. }	5	L16N10193{multip, double, multip, multiply, ..}, L18N4462, L20N6554, L22N1345, L22N1236
Division (C_{div})	{divide, division, div, /, % }	{ div, divided, divisions, U+00F7, partition, partitions, etc. }	2	L20N10457{div ,divided ,division ,U+00F7,.. }, L26N1378{div, Div, div, Div, division,.. }
Equals to (C_{eq})	{ =, total, equals, equal, equivalent }	{ equality, identical, same, exactly, contain, exact, etc. }	6	L14N7597{identical, difference, differences, equal,..}, L18N7531, L18N1850, L20N3177, L20N5535, L24N154
Calculation (C_{cal})	{formula, equation, calculation, algorithm, expression, computation }	{rewrite, sum, application, ratio, percentage, eqn, rate, etc. }	14	L11N815{equation, formula, Formula, diagram,...}, L7N7176, L8N3689, L13N2019, L15N3958

Table 3: List of concepts related to arithmetic reasoning along with their seed tokens and the count of discovered neurons in Llama2-7B. We also list the expanded tokens, promoted by the discovered neurons and their exemplar neurons. For some exemplar neurons, we also show its top-scored vocabulary tokens enclosed within braces.

5.2 RQ2: Are the discovered neurons important for eliciting the reasoning capability of LLMs?

To validate the importance of our discovered neurons, we corrupt these neurons by adding Gaussian noise to them. If these neurons are critical to LLM reasoning, the corrupted LLM should present a decrease in their performance. Specifically, for all reasoning neurons in FFN, we added noise to the neurons, changing Eqn 3 to $FFN^l(x_i^l) = \sum_{j=1}^{d_m} m_{ij}^l(v_j^l + Noise)$. As a baseline, we also corrupted the same number of random neurons for comparison. Subsequently, we run the Llama2-7B with corrupted reasoning neurons and random neurons separately. We report the few-shot CoT performance of each LLM variant on the GSM8k test set in Table 4.

We observe a substantial performance decrease of 12.29% when the discovered reasoning neurons are corrupted, in contrast to a decrease of only 5.47% observed when random neurons are corrupted. The results thus show the essential role of the discovered reasoning neurons in facilitating effective reasoning by LLMs. In addition, the performance drop when corrupting random neurons implies that some of these neurons may also play an important role (e.g., for context understanding). As shown in Table 5, these neurons reveal non-zero coefficients on average.

LLM Variant	Accuracy
No corruption	16.83%
w/ corrupted reasoning neurons	4.54%
w/ corrupted random neurons	11.37%

Table 4: Llama2-7B's performance before and after corruption of reasoning neurons vs random neurons.

5.2.1 Correlation between the reasoning performance of LLMs and the activation of their reasoning neurons

Given that the identified neurons make critical contributions to an LLM's arithmetic reasoning, a natural question is: Does an LLM's reasoning performance correlate positively with how their reasoning neurons are activated? To answer this question, we performed an experiment in the zero-shot CoT setting (Kojima et al., 2022). We specifically selected a zero-shot CoT setting for this analysis because it is unbiased due to the lack of demonstration. In our experiment, we select four zero-shot CoT prompts with varying levels of accuracy on the GSM8K test set, sourced from Yang et al. (2023). The prompts include *"Let's think step by step"*, *"Take a deep breath and work on this problem step-by-step"*, *"Break this down"*, and *"A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem"*. Their respective accuracies are 7.05%, 4.47%, 11.06%, and 5.83% in Llama2-7B. In Figure 1, we plot their accuracy along with the average coefficient of their reasoning neurons per time step during the output generation.
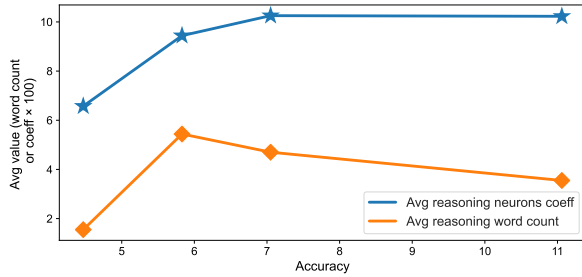
Figure 1: Correlation between prompt accuracy and the LLM's average coefficient on the discovered reasoning neurons (blue stars) and its average count of reasoning tokens (orange diamonds).

The result confirms our hypothesized positive correlation. It also reveals the potential of *predicting* an LLM's reasoning performance by examining the activation of their reasoning neurons, without needing human-annotated labels. We leave systematic explorations of this potential to the future.

Additionally, we examine if the same correlation can be observed superficially at the word level, because, if the word-level statistics present the same correlation, it could be a more convenient approach of probing into an LLM's reasoning performance than neuron activation. To respond to this question, we similarly examine the correlation between the count of reasoning tokens in the LLM generation, using a combination of the human-annotated seed tokens and the GPT-4-extracted expanded tokens listed in Table 3, and their accuracy on the GSM8k test set. Our result is presented in Figure 1. Intriguingly, we observe no positive correlation between the two factors, which thus highlights the importance of performing neuron-level analysis, as the latter offers direct insights into the functioning of LLMs that may not be visible from simply analyzing their superficial text generation.

6 Understanding Prior Observations via the Lens of Neuron Activation

In this section, we revisit the major findings from prior work and use the activation of FFN neurons in an LLM to explain them. For each research question (RQ), our analysis will be based on how each CoT prompt variant triggers different neuron activation patterns. These observations are summarized in Table 5, where the number of total or unique activated neurons is counted across the encoding steps of each CoT prompt, and the reported coefficient is an average per neuron. To show a baseline, we also report the average coefficient of randomly sampled neurons.

6.1 RQ3: Why do equations matter?

Prior works (Wang et al., 2023; Ye et al., 2023; Madaan and Yazdanbakhsh, 2022) have shown that equations play an important role in eliciting reasoning in LLMs. Looking into its activation pattern, we found the CoT prompt without equations (denoted as "w/o Equation") activates fewer reasoning neurons overall, 842 activations, compared to the CoT prompt with equations (CoT), 1119 activations. Furthermore, we observed a decrease in both the number of activated neurons for individual concepts and their corresponding average coefficients across all categories. This shows that equations play an important role in activating the reasoning neurons which are deemed to be important for arithmetic reasoning. As a result, the presence of equations can help elicit arithmetic reasoning in LLMs.

Interestingly, we also note that although there were no equations or arithmetic operators in the "w/o Equation" prompt, neurons associated with arithmetic operations (i.e., C_{add}, C_{sub}, C_{mul}, C_{div}) were still activated. This indicates that even in the absence of explicit equations in the CoT prompt, LLMs are capable of recognizing the necessity of performing arithmetic operations, which explains the 12.58% retained accuracy in Table 2.

6.2 RQ4: Why do text explanations matter?

The importance of textual explanations as found in prior work (Wang et al., 2023; Ye et al., 2023; Madaan and Yazdanbakhsh, 2022) is also consistent with our observation. We found that the CoT prompt without textual explanations activates reasoning neurons fewer times, 783 activated neurons, compared to the CoT prompt with explanation, 1119 activated neurons. Specifically, we observe a significant decrease in the activation of neurons associated with logical connectors (C_{logic}) and a slight decrease in the activation of neurons associated with arithmetic operations (particularly C_{add}). This shows the utility of textual explanations not only in activating neurons associated with logical connectors, crucial for determining the reasoning direction but also in activating neurons associated with arithmetic operations.

6.3 RQ5: Why does arithmetic diversity in exemplars matter?

Ye et al. (2023) showed that arithmetic diversity in exemplars is important for arithmetic reasoning, i.e. CoT prompts that consist of all arithmetic

Prompt Type	Logical Connectors (C_{logic})	Addition (C_{add})	Subtraction (C_{sub})	Multiplication (C_{mul})	Division (C_{div})	Equals to (C_{eq})	Calculation (C_{cal})	Reasoning Neurons (Total)	Random Neurons
CoT	(226, 35, 2.62)	(599, 15, 2.54)	(87, 2, 2.38)	(98, 4, 3.075)	(28, 2, 2.19)	(19, 4, 2.27)	(62, 6, 1.37)	(1119, 68, 2.51)	1.15
w/o Equation (RQ3)	(207, 36, 2.60)	(449, 13, 2.18)	(41, 2, 2.09)	(67, 4, 2.94)	(19, 2, 1.71)	(11, 3, 2.14)	(48, 6, 1.45)	(842, 66, 2.29)	1.53
w/o Textual Explanation (RQ4)	(85, 28, 1.50)	(450, 13, 2.94)	(86, 2, 2.34)	(73, 4, 2.95)	(51, 2, 2.46)	(11, 3, 2.68)	(27, 6, 1.03)	(783, 58, 2.62)	1.45
AddOnly (RQ5)	(286, 32, 2.55)	(651, 14, 3.04)	(97, 2, 2.62)	(173, 4, 2.24)	(13, 1, 1.66)	(40, 5, 2.24)	(90, 8, 1.25)	(1350, 66, 2.65)	1.39
MultOnly (RQ5)	(212, 25, 2.37)	(322, 11, 2.37)	(97, 2, 1.82)	(229, 5, 2.7)	(36, 1, 1.87)	(28, 4, 2.79)	(127, 10, 1.37)	(1051, 58, 2.17)	1.5
Incorrect Label (RQ6)	(221, 35, 2.60)	(601, 15, 2.52)	(97, 2, 2.41)	(103, 4, 2.86)	(28, 2, 2.10)	(24, 4, 2.26)	(65, 6, 1.35)	(1139, 68, 2.47)	1.45
OOD Label (RQ6)	(240, 41, 2.53)	(543, 15, 2.62)	(92, 2, 2.20)	(97, 4, 2.94)	(26, 2, 2.21)	(23, 4, 2.02)	(66, 7, 1.39)	(1087, 75, 2.50)	1.50

Table 5: For each prompt variant, we present (count of activated neurons, count of unique activated neurons, average coefficient) for each concept or total. We also present the average coefficient of random neurons as a baseline.

operations in their demonstrations yield better performance than the ones that do not. Our results in Table 5 indicate that the performance decline is likely caused by the bias introduced by the partial operators. We observe that the AddOnly prompt activates a higher number of C_{add} neurons (651 vs 599) and C_{mul} neurons (173 vs 98) when compared to CoT, but fewer C_{div} neurons with a lower average coefficient. Similarly, we found that MultOnly activates a significantly higher number of C_{mul} neurons when compared to the CoT prompt (229 vs 98), but significantly fewer C_{add} neurons (322 vs 599). This shows that although both AddOnly and MultOnly activate the neurons related to arithmetic reasoning, they exhibit a bias toward emphasizing specific arithmetic operations, which explains their degraded performance.

6.4 RQ6: Why does incorrect reasoning or gold label not matter?

Prior work (Wang et al., 2023; Min et al., 2022) shows that incorrect labels in the few-shot exemplars do not matter, as long as the labels come from the same distribution. Consistent with our previous findings, we observed a similar reasoning neuron activation pattern for CoT prompts with correct and incorrect labels. However, despite a 9.25% decrease in accuracy for the "OOD Label" prompt, it still exhibited a similar reasoning neuron activation pattern compared to the patterns of CoT.

To understand this phenomenon, we conducted the second analysis. In the prior work, Geva et al. (2022) found that LLMs refresh their token representations by accumulating sub-updates (Section 2.3). Therefore, two CoT prompts with similar performance presumably should reveal similar sub-updates per layer in the corresponding step, and vice versa. To validate it, we looked into the neuron activation for each prompt in the encoding steps where the labels were manipulated (e.g., the positions of "1" and "Dawson" in Table 1), as other input tokens are the same in all the three prompts. We then plotted the overlap of activated neurons per layer between CoT and "Incorrect Labels" or between CoT and "OOD Label" in Figure 4 of Appendix E. Note that here we consider all activated neurons, no matter if they are discovered as reasoning neurons or not. We observe a substantial overlap of 63.05% on average in the former case while merely 14.91% in the latter. The observation is consistent with our hypothesis, showing that the activation of FFN neurons can be used to explain the performance of CoT prompting.

The two observations (i.e., inconsistent reasoning neuron's activation pattern based on Table 5 but consistent sub-update pattern based on the overlap analysis) thus imply that the activation of reasoning neurons are *necessary* but not *sufficient* to elicit reasoning in LLMs. In fact, our qualitative analysis showed that in the case of providing OOD labels, the LLM still engages in reasoning, and their reasoning paths are similar to those prompted by correct labels (see examples in Appendix F), which explains the activation of their reasoning neurons. However, this reasoning is biased by the use of OOD tokens as variables, leading to messy variable references and an increasing amount of incorrect reasoning as the reasoning proceeds. We include a further discussion in Limitations.

7 Conclusions

Our work is among the first in applying neuron activation analysis to understanding LLMs in arithmetic reasoning. Our results offer valuable insights into the role of neurons and their utility in understanding the internal mechanism of LLMs. We thus expect this work to pave the way for future research on LLM interpretability.

Limitations

A crucial question to raise is, *does neuron activation represent all about LLM reasoning?* Despite its efficacy in explaining RQs 3-6, it is inherently limited by its focus on analyzing neurons individually without considering the interaction among neurons or other LLM components (e.g., attention modules). Consequently, a major limitation of our approach is that it may not be suitable for understanding very complex phenomena that arise from the interactions among different components of LLMs. For instance, to fully understand in-context learning within CoT, analyzing neurons in isolation may prove insufficient. Instead, as explored by Olsson et al. (2022), it requires studying attention heads and their circuits, which are sub-networks of neurons. Despite this limitation, through our study, we show that analysis of neuron activation can play an important role. Therefore, future work should study it together with other approaches such as circuit analysis (Olsson et al., 2022; Wang et al., 2022a), top-down-approach (Zou et al., 2023; Meng et al., 2022), etc. to provide a more complete picture of LLMs' inner mechanism for reasoning.

Similarly, as discussed in Section 6.4, the activation of reasoning neurons are *necessary* but not a *sufficient* to elicitate the reasoning ability of LLMs. Furthermore, their activation may only indicate the appearance of these concepts during an LLM's reasoning process, but this can be easily "faked" (e.g., prompting an LLM to produce a sequence of concept tokens pretending to be performing reasoning). As a result, the coefficient of reasoning neurons as a metric is more helpful when the prompts to LLMs are valid. Thus, it is important to exercise caution when drawing conclusions from the analysis.

Furthermore, although we employ seven concepts introduced in Section 5.1 to study arithmetic reasoning in LLMs, they may not represent the full scope of arithmetic reasoning. Hence, our study is also limited to the scope of these seven concepts.

Finally, our analyses are all based on Llama2-7B LLMs. Further exploration could verify if we can see the same observation for other LLMs as well.

Ethics Statement

We do not anticipate any severe ethical issues from using the proposed approach. We use fully open-sourced datasets and will open-source our results and dataset as well. On the other hand, we stress the positive impact of our work, as it contributes to interpreting the black box of LLMs. Forming a clear understanding of the inner mechanism of LLMs is crucial for their safe and trustworthy applications. With our investigation of neuron activation for understanding LLMs, we hope to inspire more researchers to extend the research of LLM interpretability. It is also our plan to connect neuron activation with the present societal concerns around LLM safety (e.g., analyzing an LLM's reasoning process and detecting potential vulnerabilities through their neuron activation patterns).

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2022. Analyzing transformers in embedding space. *arXiv preprint arXiv:2209.02535*.

Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2022. Linguistic correlation analysis: Discovering salient neurons in deepnlp models. *arXiv preprint arXiv:2206.13288*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscope: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought it takes two to tango.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Rafae Khan, and Jia Xu. 2022. Analyzing encoded concepts in transformer language models. *arXiv preprint arXiv:2206.13289*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022a. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022b. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Prompt for Neuron Annotation with GPT-4

To implement the `GPT4ConceptQuery` function in Algorithm 2, we query GPT-4 using the following prompt: *"A neuron in language model promotes the following set of words: $w_1, .., w_P$. Is this neuron promoting C_{name}? First, answer in Yes or No format and provide an explanation."* The function returns "Yes" when GPT-4 considers the neuron (as represented by their projected vocabulary tokens) to represent the target concept C_{name}. We additionally prompt GPT-4 to provide an explanation as it empirically motivates more precise results from GPT-4.

B Additional Details of Replicating Observations of Prior Work

Before investigating the mechanism of LLM reasoning, we first conduct experiments to replicate and validate observations made by prior work (Table 1). The experimental results based on Llama2-7B are presented in Table 2. We successfully replicated all the results of the prior work.

Although some research questions (RQs) were common in prior work, the experiment design could differ. In these cases, we opted for a more suitable or fair experiment design among them. Specifically, for RQ4, "Does textual explanation matter?", we follow the specification of Ye et al. (2023) instead of Madaan and Yazdanbakhsh (2022). Madaan and Yazdanbakhsh (2022) ablated the text and rewrites the multiple equations into a single equation to evaluate the importance of the text. We find it unfair to compare the importance of equations in the few-shot exemplar as single problem-solving steps rather than multiple steps. In our experiments, we only remove text while retaining all the equations from our original CoT instead of restructuring them into singular equations. Similarly, for RQ6, "Does correct reasoning or gold label matter?", Ye et al. (2023) proposed to manipulate only the labels of the equation. On the other hand, Wang et al. (2023) proposed to manipulate other components such as operators and textual explanations as well. We follow the specification of Ye et al. (2023) instead of Wang et al. (2023) for its simplicity and ease of analysis.

C Additional Implementation Details for Neuron Discovery (RQ1)

In Algorithm 1, we randomly select 20 examples from the GSM8k (Cobbe et al., 2021) test set as $\mathcal{E}$ and set $K = 20$. Additionally, we perform simple greedy decoding on Llama2-7B that consists of 7 billion parameters using a single NVIDIA A100 GPU for 6-7 hours to save the candidate neurons using Algorithm 1. Subsequently, we employ Algorithm 2 to identify associated neurons for each concept, with thresholds $P = 20$ and $F = 2$ where we prompt GPT-4 ~ 1300 times to obtain the reasoning neurons listed in Table 3.

D Reasoning Neurons Activation Dynamics

To better understand the activation pattern of identified reasoning neurons in Section 5.1, we plot their activation throughout an LLM's reasoning text for a randomly selected example, as shown in Figure 3 and Figure 2. Our goal is to discern the activation sites of these reasoning neurons and utilize this information to understand the role of these reasoning neurons in each reasoning step or process. To this end, we first divide the LLM's reasoning text into four sections to simplify the observation - (1) Beginning of a sentence (BOS) (2) Equations (3) Numbers (4) Other texts. The activation showed a clear pattern of activation for both neurons related to arithmetic operations and logical connections. In Figure 3, the heightened activation of arithmetic neurons, encompassing those involved in addition, subtraction, multiplication, and division, within equations is evident. Conversely, Figure 2 demonstrates increased activation of logical connection neurons at the beginning of sentences (BOS). These observations underscore the specific roles played by different neurons in the reasoning process.

E Neuron Activation Overlap between CoT with Correct Labels and Incorrect or OOD Labels

The neuron activation overlap between CoT prompt with correct labels and incorrect or OOD labels, discussed in Section 6.4 is shown in Figure 4.

F Example Predictions for CoT with OOD labels

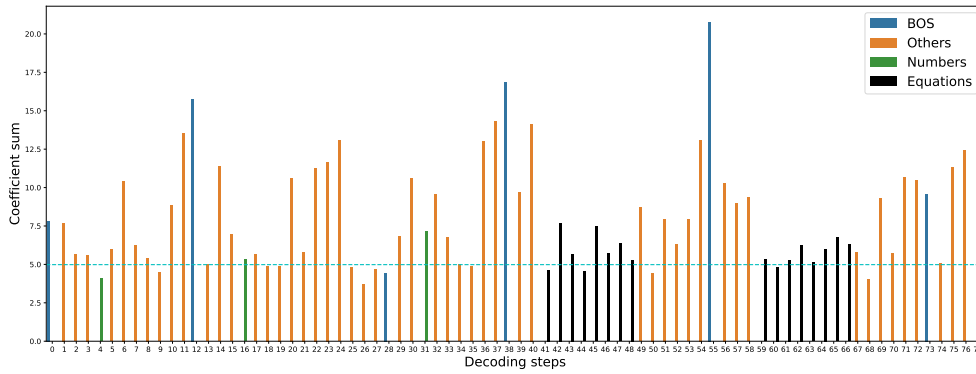We list the example prediction for CoT with OOD labels as discussed in Section 6.4 is listed in Table 6.

Figure 2: Activation pattern of logical connector neurons for a randomly sampled example. The horizontal dotted line represents the average coefficient of randomly sampled neurons for the same set of examples.
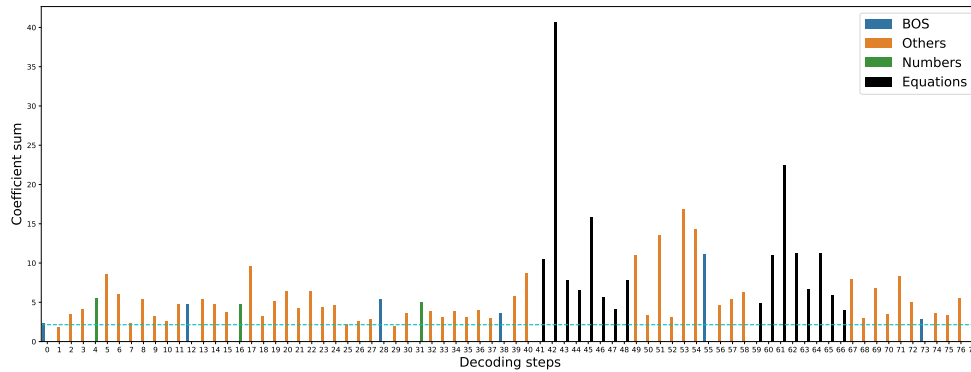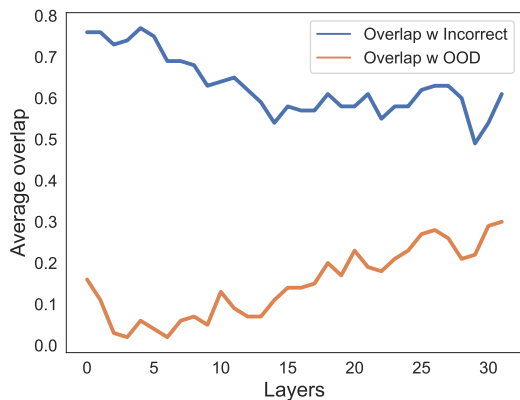

Figure 3: Activation pattern of neurons related to arithmetic neurons (add, subtraction, multiplication, division) for a randomly sampled example. The horizontal dotted line represents the average coefficient of randomly sampled neurons for the same set of examples.

Figure 4: Overlap of neuron activation when the LLM is prompted with CoT and "Incorrect Labels" or "OOD Labels".

G   Implementation Details

We use models, Llama2-7B (Touvron et al., 2023) and GPT-4 (OpenAI, 2023), and the GSM8K dataset (Cobbe et al., 2021) that are consistent with their intended use. For each experiment in our analysis (RQs 2-6), we perform simple greedy decoding on Llama2-7B which consists of 7 billion parameters using a single NVIDIA A100 GPU for 6-7 hours.

H   CoT Prompts for Reproducibility

We list all the CoT prompts used in our analysis, RQs 1-6. The CoT prompts (Correct) is listed in Table 7, CoT prompt w/o equation is listed in Table 8, CoT prompt w/o text is listed in Table 9, AddOnly prompt is listed in Table 12, MultOnly prompt is listed in Table 13, CoT prompt with incorrect labels is listed in Table 10, and CoT prompt with OOD

7185
12

Question: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?
Let's think step by step
First Janet's ducks lay 16 eggs per day.
Then she eats 3 for breakfast every day.
So she has 16 - 3 = Dylan eggs left.
Then she bakes muffins for her friends every day with 4.
So she has Dylan - 4 = Dylan eggs left.
Then she sells the remainder at the farmers' market daily for $2 per fresh duck egg.
So she has Dylan - 4 = Dylan eggs left.
So she makes $2 * Dylan = Dylan dollars.
The answer is Dylan.

Question: Two trains leave San Rafael at the same time. They begin traveling westward, both traveling for 80 miles. The next day, they travel northwards, covering 150 miles. What's the distance covered by each train in the two days?
Let's think step by step
First two trains leave San Rafael at the same time.
Then they begin traveling westward, both traveling for 80 miles.
So they travel 80 miles each.the next day, they travel northwards, covering 150 miles.
So they travel 150 miles each.
So they travel 80 + 150 = Dylan miles each.
The answer is Dylan.

Table 6: Example prediction for CoT prompts with out-of-distribution (OOD) labels

labels is listed in Table 11.

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
Let's think step by step
First there are 15 trees.
Then there were 21 trees after some more were planted.
So there must have been 21 - 15 = 6 trees.
The answer is 6.

Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
Let's think step by step
First there are 3 cars.
Then 2 more cars arrive.
So there must be 3 + 2 = 5 cars.
The answer is 5.

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
Let's think step by step
First Leah had 32 chocolates and her sister had 42 chocolates.
So in total they had 32 + 42 = 74 chocolates.
Then they ate 35 chocolates.
So there must be 74 - 35 = 39 chocolates.
The answer is 39.

Question: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
Let's think step by step
First Jason had 20 lollipops.
Then he had 12 after giving some to Denny.
So he gave Denny 20 - 12 = 8 lollipops.
The answer is 8.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
Let's think step by step
First Shawn has 5 toys.
Then he got 2 toys each from his mom and dad.
So he must have 5 + 4 = 9 toys.
The answer is 9.

Question: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
Let's think step by step
First there were 9 computers.
Then for each of 4 days, 5 more computers were added.
So 5 * 4 = 20 computers were added.
So there must be in total 9 + 20 = 29 computers.
The answer is 29.

Question: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?
Let's think step by step
First Michael started with 58 golf balls.
Then he lost 23 on Tuesday.
So he had 58 - 23 = 35 golf balls.
Then he lost 2 more on Wednesday.
So he must have 35 - 2 = 33 golf balls.
The answer is 33.

Question: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
Let's think step by step
First Olivia has 23 dollars.
Then she bought five bagels for 3 dollars each.
We know 5 bagels for 3 dollars each will be 5 * 3 = 15 dollars.
So she has 23 - 15 = 8 dollars left.
The answer is 8.

Table 7: Full prompt for **CoT** prompting for arithmetic reasoning.

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

21 - 15 = 6.

The answer is 6.

Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

3 + 2 = 5.

The answer is 5.

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

32 + 42 = 74.

74 - 35 = 39.

The answer is 39.

Question: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

20 - 12 = 8.

The answer is 8.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

5 + 4 = 9.

The answer is 9.

Question: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

5 * 4 = 20.

9 + 20 = 29.

The answer is 29.

Question: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?

58 - 23 = 35.

35 - 2 = 33.

The answer is 33.

Question: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?

5 * 3 = 15.

23 - 15 = 8.

The answer is 8.

Table 8: Full prompt for **w/o equation CoT** prompting for arithmetic reasoning.

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
Let's think step by step
First there are 15 trees.
Then there were 21 trees after some more were planted.
So there must have been 6 trees.
The answer is 6.

Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
Let's think step by step
First there are 3 cars.
Then 2 more cars arrive.
So there must be 5 cars.
The answer is 5.

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
Let's think step by step
First Leah had 32 chocolates and her sister had 42 chocolates.
So in total they had 74 chocolates.
Then they ate 35 chocolates.
So there must be 39 chocolates.
The answer is 39.

Question: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
Let's think step by step
First Jason had 20 lollipops.
Then he had 12 after giving some to Denny.
So he gave Denny 8 lollipops.
The answer is 8.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
Let's think step by step
First Shawn has 5 toys.
Then he got 2 toys each from his mom and dad.
So he must have 9 toys.
The answer is 9.

Question: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
Let's think step by step
First there were 9 computers.
Then for each of 4 days, 5 more computers were added.
So 20 computers were added.
So there must be in total 29 computers.
The answer is 29.

Question: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?
Let's think step by step
First Michael started with 58 golf balls.
Then he lost 23 on Tuesday.
So he had 35 golf balls.
Then he lost 2 more on Wednesday.
So he must have 33 golf balls.
The answer is 33.

Question: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
Let's think step by step
First Olivia has 23 dollars.
Then she bought five bagels for 3 dollars each.
We know 5 bagels for 3 dollars each will be 15 dollars.
So she has 8 dollars left.
The answer is 8.

Table 9: Full prompt for **w/o text CoT** prompting for arithmetic reasoning.

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
Let's think step by step
First there are 15 trees.
Then there were 21 trees after some more were planted.
So there must have been 21 - 15 = 1 trees.
The answer is 1.

Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
Let's think step by step
First there are 3 cars.
Then 2 more cars arrive.
So there must be 3 + 2 = 3 cars.
The answer is 3.

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
Let's think step by step
First Leah had 32 chocolates and her sister had 42 chocolates.
So in total they had 32 + 42 = 12 chocolates.
Then they ate 35 chocolates.
So there must be 12 - 35 = 45 chocolates.
The answer is 45.

Question: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
Let's think step by step
First Jason had 20 lollipops.
Then he had 12 after giving some to Denny.
So he gave Denny 20 - 12 = 2 lollipops.
The answer is 2.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
Let's think step by step
First Shawn has 5 toys.
Then he got 2 toys each from his mom and dad.
So he must have 5 + 4 = 6 toys.
The answer is 6.

Question: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
Let's think step by step
First there were 9 computers.
Then for each of 4 days, 5 more computers were added.
So 5 * 4 = 22 computers were added.
So there must be in total 9 + 22 = 23 computers.
The answer is 22.

Question: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?
Let's think step by step
First Michael started with 58 golf balls.
Then he lost 23 on Tuesday.
So he had 58 - 23 = 14 golf balls.
Then he lost 2 more on Wednesday.
So he must have 14 - 2 = 23 golf balls.
The answer is 23.

Question: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
Let's think step by step
First Olivia has 23 dollars.
Then she bought five bagels for 3 dollars each.
We know 5 bagels for 3 dollars each will be 5 * 3 = 23 dollars.
So she has 23 - 23 = 9 dollars left.
The answer is 9.

Table 10: Full prompt for **Incorrect CoT** prompting for our RQ3 experiment.

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
Let's think step by step
First there are 15 trees.
Then there were 21 trees after some more were planted.
So there must have been 21 - 15 = Dawson trees.
The answer is Dawson.

Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
Let's think step by step
First there are 3 cars.
Then 2 more cars arrive.
So there must be 3 + 2 = Russel cars.
The answer is Russel.

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
Let's think step by step
First Leah had 32 chocolates and her sister had 42 chocolates.
So in total they had 32 + 42 = Cedric chocolates.
Then they ate 35 chocolates.
So there must be Cedric - 35 = Hendricks chocolates.
The answer is Hendricks.

Question: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
Let's think step by step
First Jason had 20 lollipops.
Then he had 12 after giving some to Denny.
So he gave Denny 20 - 12 = Edison lollipops.
The answer is Edison.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
Let's think step by step
First Shawn has 5 toys.
Then he got 2 toys each from his mom and dad.
So he must have 5 + 4 = Kaylin toys.
The answer is Kaylin.

Question: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
Let's think step by step
First there were 9 computers.
Then for each of 4 days, 5 more computers were added.
So 5 * 4 = Mcgowan computers were added.
So there must be in total 9 + Mcgowan = Damarion computers.
The answer is Damarion.

Question: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?
Let's think step by step
First Michael started with 58 golf balls.
Then he lost 23 on Tuesday.
So he had 58 - 23 = Jemima golf balls.
Then he lost 2 more on Wednesday.
So he must have Jemima - 2 = Astrid golf balls.
The answer is Astrid.

Question: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
Let's think step by step
First Olivia has 23 dollars.
Then she bought five bagels for 3 dollars each.
We know 5 bagels for 3 dollars each will be 5 * 3 = Gallagher dollars.
So she has 23 - Gallagher = Baily dollars left.
The answer is Baily.

Table 11: Full prompt for Out-of-distribution (OOD) CoT prompting for arithmetic reasoning.

Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
Let's think step by step
First there are 3 cars.
Then 2 more cars arrive.
So there must be 3 + 2 = 5 cars.
The answer is 5.

Question: Paddington has 40 more goats than Washington. If Washington has 140 goats, how many goats do they have in total?
Let's think step by step
First Paddington has 40 more goats than Washington.
We know Washington has 140 goats.
So Paddington has 40 + 140 = 180 goats.
So they have 180 + 140 = 320 goats in total.
The answer is 320.

Question: Christina has 3 snakes. 1 snake is 2 feet long. Another snake is 16 inches long. The last snake is 10 inches long. How many inches are all of her snakes combined?
Let's think step by step
First Christina has 3 snakes.
Then 1 snake is 2 feet long.
We know 1 foot is 12 inches.
So 2 feet is 12 + 12 = 24 inches.
Then another snake is 16 inches long.
Then the last snake is 10 inches long.
So all of her snakes combined are 24 + 16 + 10 = 50 inches.
The answer is 50.

Question: Bush and Matt are brothers. Bush is younger than Matt by 3 years. This year Bush will be 12 years old. What will be Matt's age 10 years from now?
Let's think step by step
First Bush is younger than Matt by 3 years.
We know Bush will be 12 years old this year.
So Matt will be 12 + 3 = 15 years old this year.
Then Matt's age 10 years from now will be 15 + 10 = 25 years old.
The answer is 25.

Question: Jeremy listened to five more songs yesterday than today. Yesterday, he listened to nine songs. How many songs did Jeremy listen to in two days?
Let's think step by step
First Jeremy listened to 9 songs yesterday.
Then he listened to 5 more songs yesterday than today.
So he listened to 9 + 5 = 14 songs today.
So he listened to 9 + 14 = 23 songs in two days.
The answer is 23.

Question: Jar A has 28 marbles. Jar B has 12 more marbles than jar A. Jar C has as many marbles as jar B. How many marbles are there altogether?
Let's think step by step
First Jar A has 28 marbles.
Then Jar B has 12 more marbles than jar A.
So Jar B has 28 + 12 = 40 marbles.
Then Jar C has as many marbles as jar B.
So Jar C has 40 marbles.
So there are 28 + 40 + 40 = 108 marbles altogether.
The answer is 108.

Question: Marion received 20 more turtles than Mia at the animal rescue center. If Mia received 40 turtles, how many turtles did they receive together?
Let's think step by step
First Marion recieved 20 more turtles than Mia.
We know Mia received 40 turtles.
So Marion received 20 + 40 = 60 turtles.
So together they received 60 + 40 = 100 turtles.
The answer is 100.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
Let's think step by step
First Shawn has 5 toys.
Then he got 2 toys each from his mom and dad.
So he must have 5 + 4 = 9 toys.
The answer is 9.

Table 12: Full prompt for **AddOnly CoT** prompting for arithmetic reasoning.

Question: Super Clean Car Wash Company cleans 80 cars per day. They make $5 per car washed. How much money will they make in 5 days?
Let's think step by step
First Super Clean Car Wash Company cleans 80 cars per day.
Then they make 5 dollars per car washed.
So they make 80 x 5 = 400 dollars per day.
So they make 400 x 5 = 2000 dollars in 5 days.
The answer is 2000.

Question: A farmer has 5 cows. Each cow has 4 legs. How many legs do the cows have in total?
Let's think step by step
First a farmer has 5 cows.
Then each cow has 4 legs.
So the cows have 5 x 4 = 20 legs in total.
The answer is 20.

Question: Sam watches two movies each day. Each movie is 2 hours long. How many minutes does Sam spend watching movies in 5 days?
Let's think step by step
First Sam watches two movies each day.
Then each movie is 2 hours long.
We know 1 hour is 60 minutes.
So 2 hours is 60 x 2 = 120 minutes.
So Sam spends 120 x 2 = 240 minutes watching movies each day.
So Sam spends 240 x 5 = 1200 minutes watching movies in 5 days.
The answer is 1200.

Question: Carla has 3 bags. Each bag has 5 apples. How many apples does Carla have in total?
Let's think step by step
First Carla has 3 bags.
Then each bag has 5 apples.
So Carla has 3 x 5 = 15 apples in total.
The answer is 15.

Question: James takes 20 units per semester at community college. If each unit costs $50 how much does he pay for 2 semesters?
Let's think step by step
First James takes 20 units per semester at community college.
Then each unit costs 50 dollars.
So he pays 20 x 50 = 1000 dollars per semester.
So he pays 1000 x 2 = 2000 dollars for 2 semesters.
The answer is 2000.

Question: In a jar that has 50 ants, the number of ants in the jar doubles each hour. How many ants will be in the jar after 5 hours?
Let's think step by step
First there are 50 ants in the jar.
Then the number of ants in the jar doubles each hour.
So there will be 50 x 2 = 100 ants in the jar after 1 hour.
So there will be 100 x 2 = 200 ants in the jar after 2 hours.
So there will be 200 x 2 = 400 ants in the jar after 3 hours.
So there will be 400 x 2 = 800 ants in the jar after 4 hours.
So there will be 800 x 2 = 1600 ants in the jar after 5 hours.
The answer is 1600.

Question: Mark loves to see shows in theaters. He decided to visit the theater at least once a week. One performance lasts 3 hours. The price of the ticket depends on the time spent in the theater and stands at $5 for each hour. How much will Mark spend on visits to the theater in 6 weeks?
Let's think step by step
First Mark decided to visit the theater at least once a week.
Then one performance lasts 3 hours.
We know the price of the ticket depends on the time spent in the theater and stands at 5 dollars for each hour.
So the price of the ticket for one performance is 5 x 3 = 15 dollars.
So Mark will spend 15 x 6 = 90 dollars on visits to the theater in 6 weeks.
The answer is 90.

Question: A sixty bulb watt uses 60 watts of power each day. If Allyn has 40 such bulbs in his house and pays an electricity bill of twenty cents per power watt used, calculate Allyn's total monthly expenses on electricity in June.
Let's think step by step
First a sixty bulb watt uses 60 watts of power each day.
Then Allyn has 40 such bulbs in his house.
So Allyn has 40 x 60 = 2400 watts of power each day.
Then Allyn pays an electricity bill of twenty cents per power watt used.
So Allyn pays 2400 x 0.2 = 480 dollars per day.
So Allyn pays 480 x 30 = 14400 dollars per month.
The answer is 14400.

Table 13: Full prompt for **MultOnly CoT** prompting for arithmetic reasoning.