Activation Differential Analysis for Enhancing Chain-of-thought Reasoning

Anonymous ACL submission

Abstract

Despite the impressive chain-of-thought (CoT) reasoning ability of large language models (LLMs), its underlying mechanisms remains unclear. In this paper, we explore the inner workings of LLM's CoT ability via the lens of neurons in the feed-forward layers. We propose an efficient method to identify reasoning-critical neurons by analyzing their activation patterns under reasoning chains of varying quality. Based on it, we devise a rather simple intervention method that directly stimulates these reasoning-critical neurons, to guide the generation of high-quality reasoning chains. Extended experiments validate the effectiveness of our method and demonstrate the critical role these identified neurons play in CoT reasoning. Our code and data will be publicly available.

1 Introduction

Through the chain-of-thought (CoT) prompting strategy (Wei et al., 2022; Merrill and Sabharwal, 2024), large language models (LLMs) can arrive at correct answers through a step-by-step reasoning paradigm. However, LLMs often generate text with obvious mistakes, raising doubts about their ability to robustly process reasoning chains (Turpin et al., 2023). Therefore, understanding LLMs reasoning mechanisms is important to improve their reasoning accuracy and efficiency.

A surge of work has been conducted to explore techniques to improve reasoning accuracy and efficiency. Previous studies have predominantly focused on optimizing external components of CoT (Fu et al., 2023; Wang et al., 2023a; Tang et al., 2023; Jin et al., 2024), such as prompt engineering and symbolic representations (Madaan and Yazdanbakhsh, 2022; Ye et al., 2023). While these approaches provide valuable external insights into the factors that enhance CoT performance, they fall short of offering an internal explanation for the quality of the model's outputs.

To address this gap, researchers have attempted to provide mechanistic explanations for the model's CoT reasoning abilities. Existing work can be roughly categorized into module-level and neuron-level interpretation methods. Concretely, the module-level methods generally leverage causal tracing (Meng et al., 2022, 2023) and circuit construction (Hanna et al., 2023; Yao et al., 2024) to identify and analyze key modules involved in the model's CoT reasoning process. However, due to the higher cost of estimating all the components within LLMs, these methods can not be used for more fine-grained analysis, attention heads and neurons. In contrast, neuron-level methods aim to identify important neurons in the model by analyzing their activation values in the feed-forward network (FFN) (Stolfo et al., 2023; Yu and Ananiadou, 2024b,a) or attention heads (Wang et al., 2023b; Li et al., 2023; Yeh et al., 2024). However, the large scale of the neurons and their great randomness in activation values, also increase the difficulty in accurately estimating their contributions.

In this paper, we identify reasoning-critical neurons by leveraging the activation differences of FFN neurons across reasoning chains of varying quality. Our motivation is that by modulating their activation strengths, we can directly enhance the model performance on downstream tasks. Concretely, we propose an efficient approach to investigate the inner workings of LLMs' reasoning abilities through the lens of neurons in the feed-forward layers. We first construct a contrastive dataset of varying reasoning trajectories using the MATH benchmark's training set. Leveraging the dataset, we analyze the neurons activation patterns under reasoning chains of varying quality. Specifically, we quantify the disparity in neuron activations by computing the ratio of their activation values between high- and low-quality chains, then apply a threshold to select neurons exhibiting significant activation differences. As shown in Figure 4a, these neurons consistently demonstrate stronger activation during correct reasoning chains. Then, we modulate the activation strengths of these neurons to alter the quality of generated CoT chains.

Experimental results demonstrate the effectiveness of our method across all subdomains of the MATH benchmark, leading to 2.4% relative improvement on average.

2 Preliminary

Currently, most LLMs are built upon an autoregressive Transformer architecture (Vaswani et al., 2017), in which the core components are the multi-head self-attention (MHA) and the feed-forward network (FFN). Given the MHA output $\boldsymbol{h}_i^l$ at layer i, the FFN output can be expressed as follows:

$$FFN(\boldsymbol{h}_i^l) = \boldsymbol{V}^l f(\boldsymbol{K}^l \boldsymbol{h}_i^l) \tag{1}$$

where $\boldsymbol{K}^l \in \mathbb{R}^{N \times d}$,$\boldsymbol{V}^l \in \mathbb{R}^{d \times N}$ represent two linear layers, and f denotes the non-linear activation function. In this paper, we define a neuron as a specific scalar parameter in the weight matrix $\boldsymbol{V}^l$.

In this paper, we study how to identify the activation coefficients of key neurons within the LLM, and how to improve the CoT reasoning ability by intervening these neurons.

3 Methodology

3.1 Contrastive Dataset Construction

To identify neurons that significantly influence the quality of CoT, we first construct a contrastive dataset of high-quality and low-quality CoT reasoning trajectories using the MATH benchmark's training set, which covers seven mathematical subdomains to diversity in the thematic content of reasoning tasks. For each problem, we generate multiple CoT trajectories through controlled sampling, which ensures that each problem contains 5 to 10 different model outputs. Then we classify them into quality categories based on solution quality. We perform initial classification based on answer correctness, then we conduct manual verification, ultimately obtaining a contrastive dataset that encompasses both high- and low-quality CoT instances. High-quality CoT demonstrates both correct final answers and logically consistent reasoning steps, while low-quality CoT contains either incorrect answers or fundamentally flawed reasoning paths. The final dataset comprises 4,900 meticulously constructed CoT pairs for neuron identification.

3.2 CoT Key Neurons Identification

Neuron Contribution Estimation. Based on our contrastive dataset, we analyze the internal activation differences in the model under different quality CoTs, to estimate the contribution of each neuron on generating high-quality CoTs. Specifically, we feed the LLM with CoT trajectories. For the j-th neuron in the i-th layer, we first compute the average activation strength when processing the CoT trajectories. We define $m_{ij}^{(+)}$ as the average activation strength value for the high-quality CoT trajectories and $m_{ij}^{(-)}$ for the low-quality CoT trajectories. Given the varying average activation values of neurons across different layers, defining an appropriate significance threshold is challenging. Therefore, we consider using ratio-based differentiation $r_{ij} = m_{ij}^{(+)}/m_{ij}^{(-)}$ rather than absolute difference metrics to quantify the neuronal variance.

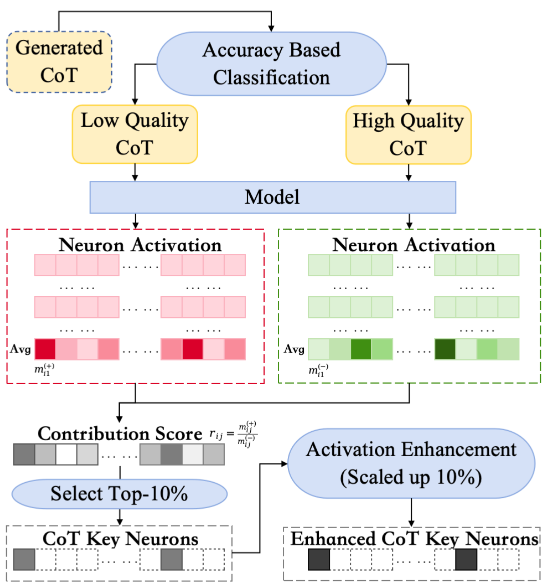


Figure 1: CoT key neuron identification and intervention based on FFN neurons activation difference.

CoT Key Neurons Selection. Our identification protocol employs a cascaded filtering approach: first, we select neurons in the top 10% of the $\{r_{ij}\}$ distribution, then we impose a predefined threshold to further filter neurons with significant differences. If the difference measure r_{ij} of a neuron exceeds this threshold, we consider that neuron to be related to the quality of the LLM's CoT. We present this step in Algorithm 1 in Appendix.

Interventing Neurons for Improving CoT Reasoning. We next validate whether our method suc-

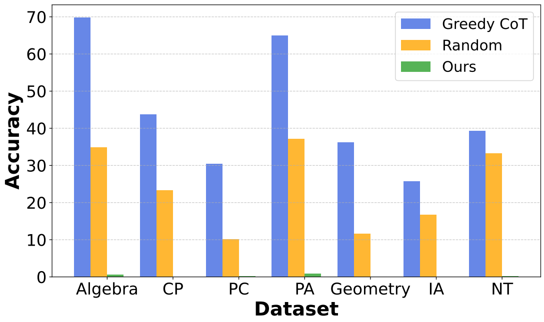
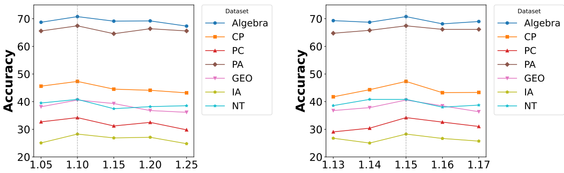


Figure 2: Impact of perturbing neuron activation values on the reasoning task accuracy of LLaMA-3.2 (3B).



(a) Scaling scalar

(b) Threshold

Figure 3: Impact of selection threshold and scaling scalar on the reasoning accuracy of LLaMA-3.2 (3B).

cessfully identifies reasoning neurons. We begin by conducting a neuron coefficient enhancement experiment, where we amplify the coefficients of the identified neurons and observe the resulting performance changes on downstream tasks. Following this, we perform a neuron coefficient interference experiment, in which we set the coefficients of the identified neurons to zero and examine the impact on performance in downstream tasks.

4 Experiments

4.1 Main Results

Here, we present our experimental findings, our experimental setup is presented in Appendix. We first identify a set of critical neurons through our proposed method, which selects neurons exhibiting significantly higher activation strength under high-quality reasoning chains compared to low-quality instances. We then conduct enhancement experiments by amplifying the activation values of these neurons by 1.1 during mathematical reasoning tasks. For comparison, we evaluate three baseline conditions with equivalent quantities of neurons, detailed descriptions of these methods are provided in Appendix. The main results are presented in Table 2, we observe that the enhancement of our identified differential neurons yields consistent accuracy improvements across all MATH subdatasets, with average gains of 2.4% compared to greedy CoT. This performance advantage suggests that our methodology effectively captures neurons specifically involved in high-quality reasoning processes, potentially responsible for steering LLM to generate high quality reasoning chains.

To further investigate the causal relationship between these neurons and reasoning capability, we conduct interference experiments through activation suppression. We observe that complete deactivation of these neurons result in catastrophic failure on solving mathematical problems. In contrast, random deactivation of equivalent numbers of neurons only causes relatively marginal performance decreases. This sharp contrast in task sensitivity confirms that the identified neurons are crucial for maintaining mathematical reasoning capabilities.

4.2 Further Analysis

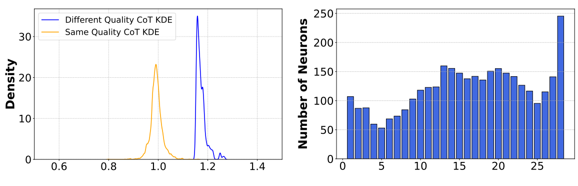
Ablation study. Here, we conduct experiments to investigate the influence of two hyper-parameters in our method. We first examine the impact of the threshold used to select neurons. The results are shown in Figure 3a, as the selection threshold increases, neurons associated with CoT quality are identified, leading to a gradual improvement in the pruned model's accuracy on mathematical reasoning tasks. However, further elevation of the selection threshold may result in the exclusion of critical neurons, causing a decline in the model's task performance. We then set the selection threshold to 1.15, exploring the impact of varying scaling factors. As shown in Figure 3b, increasing the scaling factor enhances the pruned model's reasoning ability. However, as the scaling factor continues to grow, the model's performance begins to decline, which is likely attributed to the model's sensitivity to the activation coefficients.

Activation pattern under varying quality CoTs. As shown in Figure 4a, when comparing activation patterns between high-quality and low-quality CoTs, we observe distinct distribution characteristics. Neurons activated under different quality CoT samples exhibit a pronounced ratio peak around 1.16, while those from same-quality CoT samples reveal no significant ratio differences. This validates our method's capability to isolate reasoning-critical neurons through cross-quality comparisons.

Neuron distribution across layers. Figure 4b presents the distribution of average identified neu-

Models	Method	MATH							
		Algebra	CP	PC	PA	Geometry	IA	NT	Avg.
LLaMA 3.2 3B IT	Greedy CoT	69.75	43.68	30.40	65.00	36.15	25.8	39.32	47.71
	Top-activation	67.96	43.68	32.50	63.72	37.80	23.40	**42.32**	47.34
	MathNeuro	67.96	44.53	29.00	65.23	38.47	26.50	39.70	47.64
	Random	69.15	43.00	30.20	65.50	36.15	26.15	36.70	47.35
	Ours	**70.77**	**47.32**	**33.65**	**67.44**	**40.59**	**28.27**	40.82	**50.11**
LLaMA 3.1 8B IT	Greedy CoT	67.80	41.32	31.16	67.90	36.36	26.90	**42.69**	48.20
	Top-activation	66.27	42.82	31.73	67.90	35.70	26.76	41.57	47.83
	MathNeuro	68.82	41.97	31.50	68.00	36.36	27.34	42.50	48.61
	Random	66.53	42.50	30.85	66.83	35.92	26.50	40.43	47.43
	Ours	**69.07**	**46.04**	**33.26**	**69.88**	**40.59**	**28.27**	42.32	**50.13**
LLaMA 3.2 1B IT	Greedy CoT	44.52	23.76	17.20	41.74	22.83	12.74	21.16	28.74
	Top-activation	42.30	24.10	16.80	41.00	22.26	12.63	19.10	27.78
	MathNeuro	44.85	23.76	14.50	42.79	24.52	12.63	21.9	28.93
	Random	45.19	23.80	17.00	40.50	21.80	13.00	20.78	28.57
	Ours	**47.32**	**26.33**	**19.12**	**44.3**	**26.84**	**14.13**	**24.34**	**31.28**
Qwen Math 2.5B IT	Greedy CoT	91.42	68.31	60.99	84.88	64.06	59.79	78.65	75.05
	Top-activation	91.75	68.52	63.47	84.88	63.42	58.63	76.02	74.87
	MathNeuro	91.68	69.59	61.76	84.76	64.75	61.29	78.15	75.58
	Random	91.50	68.31	61.18	84.65	63.42	59.55	79.13	75.00
	Ours	**92.77**	**70.88**	**63.67**	**86.27**	**65.96**	**61.64**	**80.90**	**76.91**

Table 1: Experimental results on MATH dataset. PC and PA denote *Precalculus* and *Prealgebra*, respectively. Avg. is the average value of all categories. The best are denoted in bold and the second-best are underlined.



(a) Activation difference (b) Neuron distribution

Figure 4: Distribution of activation strength difference and identified reasoning neurons across layers.

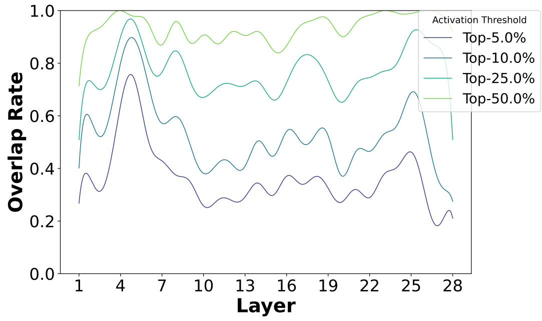


Figure 5: Overlap between the identified neurons and the top-activated neurons across layers.

rons across model layers. Reasoning-critical neurons predominantly cluster in middle-to-high layers, with the final layer containing most identified neurons. This distribution aligns with prior findings about transformer architectures, where middle layers encode task-solving information while final layers specialize in answer generation. The high concentration in later layers suggests these neurons serve as final-stage quality controllers that integrate intermediate reasoning states into coherent outputs.

Overlap between the identified neurons and the top-activated neurons. Figure 5 illustrates the overlap rates between the neurons identified by our method and the top $5\% - 50\%$ activated neurons across different layers, revealing a U-shaped pattern. It indicates that critical neurons for reasoning quality are not consistently among the most highly activated neurons, particularly in middle layers. It aligns with our experimental findings that scaling the activation values of neurons with significant activation differences across reasoning qualities within the top-activated group yields weaker performance improvements compared to scaling all neurons with significant activation differences.

5 Conclusion

In this work, we investigate the internal activation patterns of models when generating Chain-of-Thought (CoT) of varying quality. Specifically, we first construct a contrastive dataset comprising correct and incorrect reasoning chains, then we propose an effective method to identify reasoning-critical neurons based on activation disparities. Through further experiments, we demonstrate that modulating the activation strengths of these neurons can enhance the model's reasoning performance on downstream tasks.

Limitations

Our study has several limitations. First, our analysis experiments are primarily conducted on the LLaMA-3.2-3B architecture. Since neural sensitivity to interventions varies significantly across model families and scales, some conclusions of our analysis results may not generalize to other LLMs. Second, while we focus on FFN layers due to their established role in knowledge representation (Dai et al., 2022), LLMs' reasoning ability comes from complex interactions between multiple components, so a complete mechanistic understanding requires future investigation into more components in LLMs like attention layers. Finally, although our contrastive dataset for identifying reasoning neurons is effective, we have not systematically explored optimal dataset characteristics for neuron identification, we plan to explore these in our future work.

References

Bryan R. Christ, Zack Gottesman, Jonathan Kropko, and Thomas Hartvigsen. 2024. Math neurosurgery: Isolating language models' math reasoning abilities using only forward passes. *CoRR*, abs/2410.16930.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Mor Geva, Avi Caciularu, Kevin Ro 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 30–45. Association for Computational Linguistics.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1830–1842. Association for Computational Linguistics.

Zongxia Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. 2023. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *CoRR*, abs/2307.05052.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *CoRR*, abs/2209.07686.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

William Merrill and Ashish Sabharwal. 2024. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

MetaAI. 2024a. Introducing Llama 3.1: Our most capable models to date.

MetaAI. 2024b. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7035–7052. Association for Computational Linguistics.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. *CoRR*, abs/2305.14825.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2717–2739. Association for Computational Linguistics.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9840–9855. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *CoRR*, abs/2405.17969.

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, Toronto, Canada, July 9-14, 2023*, pages 4469–4484. Association for Computational Linguistics.

Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda B. Viégas, and Martin Wattenberg. 2024. Attentionviz: A global view of transformer attention. *IEEE Trans. Vis. Comput. Graph.*, 30(1):262–272.

Zeping Yu and Sophia Ananiadou. 2024a. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3293–3306. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024b. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3267–3280. Association for Computational Linguistics.

A Reasoning Neuron Collection Algorithm

We present our proposed neuron collection method in Algorithm 1

Algorithm 1 Reasoning Neuron Collection

Input: Correct solution examples $\mathcal{E}_1$, incorrect solution examples $\mathcal{E}_2$, selection ratio threshold θ, the target LLM
Output: A set of candidate neurons $\mathcal{N}$.
Initialize $\mathcal{N} \leftarrow \{\}, M_{ij}^{(+)} \leftarrow 0, M_{ij}^{(-)} \leftarrow 0$
for each example in $\mathcal{E}_1$:
  for each layer $i = 1, ..., m$:
    for each neuron $j = 1, ..., n$:
      $\hat{a}_{ij} \leftarrow \texttt{AvgL2Norm}(\{a_{ij}^k\}_{k=1}^N, k)$
      $M_{ij}^{(+)} \leftarrow M_{ij}^{(+)} + \hat{a}_{ij}$
for each example in $\mathcal{E}_2$:
  for each layer $i = 1, ..., m$:
    for each neuron $j = 1, ..., n$:
      $\hat{a}_{ij} \leftarrow \texttt{AvgL2Norm}(\{a_{ij}^k\}_{k=1}^N, k)$
      $M_{ij}^{(-)} \leftarrow M_{ij}^{(-)} + \hat{a}_{ij}$
for each layer $l = 1, ..., L$:
  for each neuron $j = 1, ..., n$:
    $m_{ij}^{(+)} \leftarrow \texttt{Avg}(M_{ij}^{(+)}, size(\mathcal{E}_1))$
    $m_{ij}^{(-)} \leftarrow \texttt{Avg}(M_{ij}^{(+)}, size(\mathcal{E}_2))$
    $\{r_{ij}\} \leftarrow \texttt{FindLargest}(m_{ij}^{(+)}/m_{ij}^{(-)}, \theta)$
    $\mathcal{N} \leftarrow \mathcal{N} \cup \{v_{ij} | r_{ij} \in \{r_{ij}\}\}$

B Experimental Setup

Models. We conduct our primary experiments on LLaMA 3.2 3B Instruct (MetaAI, 2024b), a state-of-the-art language model specifically fine-tuned for instruction-following and reasoning tasks. LLaMA 3.2 3B Instruct is known for its robust performance in complex reasoning scenarios, particularly in mathematical and logical problem-solving, making it an ideal candidate for our study on CoT reasoning.To ensure the generalizability of our approach, we also evaluate our method on models of varying scales and architectures, including LLaMA 3.2 1B (MetaAI, 2024b) Instruct ,LLaMA 3.1 8B Instruct (MetaAI, 2024a) and Qwen Math 2.5 Instruct. This multi-model setup allows us to validate the applicability of our method across different configurations.

Dataset. Our evaluation is conducted on the test sets of the MATH benchmark (Hendrycks et al., 2021), a widely recognized dataset designed to assess the mathematical reasoning and problem-solving capabilities of large language models. The MATH dataset comprises a collection of challenging competition-level mathematical problems, typically sourced from middle and high school math competitions such as AMC and AIME. These problems span a broad range of mathematical domains and are carefully curated to test reasoning skills. The dataset is divided into seven categories: Algebra, Counting and Probability, Precalculus, Prealgebra, Geometry, Intermediate Algebra, and Number Theory, providing a comprehensive benchmark for our study. The details of the datasets is shown in Table 2.

Category	Train	Dev/Test
Algebra	1744	1187
CP	771	474
Precalculus	746	546
Prealgebra	1205	871
Geometry	870	479
IA	1295	903
NT	869	540

Table 2: Statistics of the MATH datasets. CP, IA, and NT denote *Counting and Probability*, *Intermediate Algebra*, and *Number Theory*, respectively.

C Details of Main Experiments Baselines

• *Top Activated Neurons.* Many existing methods directly identify important neurons through saliency scores (Geva et al., 2022; Sun et al., 2024). Inspired by prior work, we select the top $K\%$ of neurons with the highest average activation values under positive CoT conditions as important neurons. This approach provides a computationally efficient baseline for neuron identification.

• *MathNeuro.* MathNeuron (Christ et al., 2024) identifies important parameters in LLMs by isolating math-specific parameters and improves downstream task performance through parameter scaling and pruning. We adapt this method to a neuron-level version by identifying neurons that are activated under positive CoT but not under negative CoT conditions. We use its default implementation for our pruning experiments.

• *Random Selection.* As a control baseline, we randomly select the same number of neurons to compare against the other methods. This baseline serves as a reference for different methods.

D Domain-Specific Neuron Analysis

To investigate relationships between selected neurons from different mathematical reasoning datasets, we perform set operations on neurons filtered by seven domain-specific contrastive datasets.

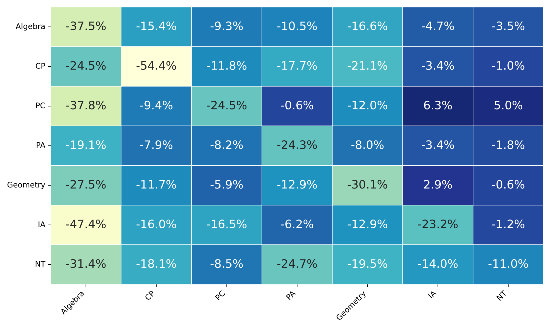


Figure 6: Pertubation result across different domain-specific neurons.

By computing the complement of each dataset-specific neuron set against the union of all other domain sets, we identify unique neurons exclusively associated with individual mathematical domains, which we term domain-specific neurons. The quantitative distribution of these neurons across domains is presented in Table 3. We further conduct intervention experiments to examine the impact of these specific neurons, the results are presented in Figure 6, we observe that suppressing activation values of domain-specific neurons in domain A causes disproportionately larger accuracy degradation on Domain A's evaluation set compared to other domains. This suggests that beyond general mathematical reasoning neurons, activation patterns of neurons tied to particular mathematical subfields also contribute to LLM's CoT reasoning quality.

Algebra	**CP**	**PC**	**PA**	**Geometry**	**IA**	**NT**
1,580	1,071	2,880	604	4,246	492	278

Table 3: The number of neurons across different domains.

Inspired by prior work (Geva et al., 2022), we further project these neurons to vocabulary space via unembedding matrices. As exemplified in Table 4, we observe that some domain-specific neurons exhibit semantic associations with their corresponding mathematical domains, which provides additional evidence for our hypothesis that domain-specific neurons constitute modular knowledge units specialized for distinct reasoning contexts.

Category	neuron	Top tokens
Geometry	f_{23}^{431}	Vol, vol, volume, Vol, vol
	f_{26}^{1727}	sphere, spherical, spheres, Sphere, Sphere
	f_{26}^{1806}	radius, radius, Radius, Radius, _radius
Algebra	f_{18}^{7100}	vectors, vector, Vector, vector, direction
	f_{24}^{4347}	Distance, distance, Distance, distances, distance
	f_{19}^{391}	projection, projections, blitz, project, optimal
NT	f_{23}^{2802}	Ninth, Nine, Sep, XIII, IX
	f_{25}^{5198}	567, 42, 345, 678, 876
	f_{26}^{937}	third, Third, Third, -three, third
CP	f_{14}^{1452}	sum, total, sum, .sum, total
	f_{19}^{2920}	more, more, 更多, More, MORE
	f_{19}^{4955}	percentage, percentages, percent, Percentage, Percent

Table 4: List of domains related to math reasoning along with their relative neurons and neurons' corresponding top tokens in Llama 3.2-3B Instruct.