

Discovering Decoupled Function Modules in Large Language Models

Yanke Yu¹, Jin Li¹, Ying Sun^{1*}, Ping Li², Zhefeng Wang², Yi Zheng²

¹Thrust of AI, Hong Kong University of Science and Technology (Guangzhou)

²Huawei Technologies Ltd.

rankyurky@gmail.com, jli740@connect.hkust-gz.edu.cn, yings@hkust-gz.edu.cn
{liping61, wangzhefeng, zhengyi29}@huawei.com

Abstract

Understanding the internal functional organization of Large Language Models (LLMs) is crucial for improving their trustworthiness and performance. However, how LLMs organize different functions into modules remains highly unexplored. To bridge this gap, we formulate a function module discovery problem and propose an Unsupervised LLM Cross-layer Module Discovery (ULCMOD) framework that simultaneously disentangles the large set of neurons in the entire LLM into modules while discovering the topics of input samples related to these modules. Our framework introduces a novel objective function and an efficient Iterative Decoupling (IterD) algorithm. Extensive experiments show that our method discovers high-quality, disentangled modules that capture more meaningful semantic information and achieve superior performance in various downstream tasks. Moreover, our qualitative analysis reveals that the discovered modules show function comprehensiveness, function hierarchy, and clear function spatial arrangement within LLMs. Our work provides a novel tool for interpreting LLMs' function modules, filling a critical gap in LLMs' interpretability research.

Code/Appendix — <https://github.com/rank-Yu/llm-modules>

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable human-like capabilities, such as mathematical problem-solving, coding, and information seeking (Zhao et al. 2023b; Gong and Sun 2024; Xin et al. 2025). However, the internal functional mechanisms of how LLMs perform complicated tasks remain unclear. A deeper understanding is crucial to improve trustworthiness, support diagnosis of model performance, and benefit potential improvements in model capabilities (Huang et al. 2025; Xu, Jain, and Kankanhalli 2024; Yao et al. 2023; Wang et al. 2024; Guo et al. 2025).

Recent studies have examined LLM internal patterns from various perspectives, such as neuron activation patterns (Bills et al. 2023; Voita, Ferrando, and Nalmpantis 2023), how models represent abstract concepts (Bricken et al. 2023; Templeton et al. 2024; Lindsey et al. 2025), or to analyze computational circuits (Hanna, Liu, and Variengien 2023; Conmy

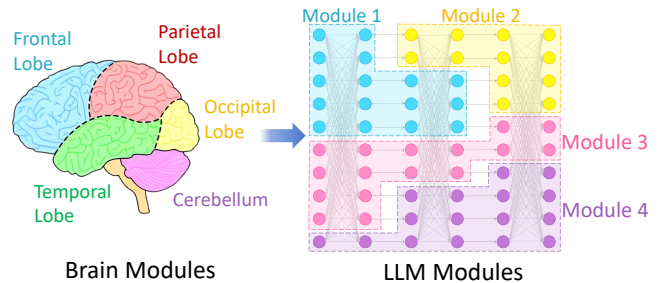


Figure 1: Illustrations of brain modules vs. LLM modules.

et al. 2023b; Ameisen et al. 2025). Interestingly, some studies have shown evidence for diversified activation patterns for different functions in LLMs. For example, Wang et al. (2022b); Panigrahi et al. (2023); Zhao et al. (2023a); Tang et al. (2024); Zhao et al. (2024b) analyzed activated neurons for inputs with different predefined topic labels and discovered distinct distributions among these neurons. Zhang et al. (2023); Xiao et al. (2024) found that specific neurons are responsible for distinct functions. Templeton et al. (2024) used Sparse AutoEncoder to learn dictionary features for LLM activations, finding that the embeddings of the dictionary features form clusters that align with their semantics.

Indeed, neuroscience research suggests that the human brain contains highly specialized and decoupled function modules (Meunier, Lambiotte, and Bullmore 2010; Bullmore and Sporns 2009; Kandel et al. 2013) shown in Fig. 1 (left). These modules facilitate parallel processing, reduce interference between cognitive functions, and allow for greater evolutionary adaptability. The activation pattern differences across functions can provide meaningful insights into a similar modularity phenomenon in LLMs, as shown in Fig. 1 (right). However, existing works only identify significantly overlapping neuron sets activated by predefined functions (Xiao et al. 2024). How different functions are formulated by basic modules and organized within LLMs remains highly unexplored. To our knowledge, it still lacks an effective method to discover truly decoupled function modules.

To bridge this gap, we formulate a **function module discovery** problem, aiming to discover decoupled neuron sets that have dense co-activation performance on a specific set of samples with a shared but unknown topic. We propose an Unsupervised LLM Cross-layer Module Discovery (ULCMOD)

*Corresponding author.

framework that simultaneously disentangles the large set of neurons in the entire LLM into modules while discovering the topics of input samples related to these modules. Specifically, we formulate a dual row-column partition problem on the LLM activation matrix and introduce a novel objective function that optimizes intra-module activation density while balancing the sizes of modules. To address the combinatorially vast search space, we develop an Iterative Decoupling (IterD) algorithm that alternates between adjusting neuron and sample partitions to optimize the objective function.

We conduct extensive experiments on the Qwen2.5 LLM family. The experimental results demonstrate our method outperforms existing clustering algorithms in identifying high-quality disentangled modules. These function modules capture more meaningful semantic information and achieve superior performance in various downstream tasks. Furthermore, our qualitative analysis reveals that the discovered modules demonstrate function comprehensiveness, function hierarchy, and clear function spatial arrangement within LLMs.

Our main contributions can be summarized as follows:

1. Formulate the function module discovery problem. To our knowledge, we are among the first to explore discovering decoupled function modules in LLMs, filling a critical blank in LLM interpretability research.
2. Introduce a novel ULCMOD framework with a new objective function and an IterD algorithm, effectively identifying high-quality and structurally sound function modules.
3. Provide extensive quantitative and qualitative analysis to validate function module discovery performance and offer insights into patterns of function organization in LLMs.

2 Related Work

Here, we provide related papers for a better understanding.

LLM Interpretability Recently, the interpretability of LLMs has gradually drawn attention of the research community (Sun et al. 2021; Zhao et al. 2024a; Ji et al. 2025; Sun, Zhu, and Xiong 2025), with several paradigms emerging to deconstruct their internal mechanisms, including *causal tracing*, *computational circuits*, and *Sparse Autoencoders (SAEs)*. One paradigm, *causal tracing*, seeks to understand information flow by observing how interventions on specific parameters or activations affect the final output (Meng et al. 2022; Geva et al. 2023; Stolfo, Belinkov, and Sachan 2023; Zhang and Nanda 2023). These methods typically locate critical model components by modifying their states and observing the change in the final prediction. These methods could help understand the importance of different components in LLM. Another paradigm focuses on constructing *computational circuits* from input to output that underlie specific model functions and behaviors, treating components like attention heads and FFN layers as fundamental units (Elhage et al. 2021; Wang et al. 2022a; Olsson et al. 2022; Hanna, Liu, and Variengien 2023; Conmy et al. 2023a,b; Ameisen et al. 2025). These methods could explore the function of different attention heads or construct the entire computational circuits. More recently, methods using *Sparse Autoencoders (SAEs)* have gained traction for learning more interpretable dictionary features from LLM activations (Bricken et al. 2023;

Templeton et al. 2024; Lindsey et al. 2025). Although these approaches have significantly advanced LLM interpretability, the analysis of decoupled function modules remains a largely unexplored area. This paper addresses this gap by introducing a novel framework for identifying and analyzing these modules, thereby providing a new perspective on the internal functional mechanisms of LLMs.

LLM Functional Activation Pattern Analysis Recent research of LLMs has revealed diversified activation patterns corresponding to their various functions. They analyzed activated neurons for inputs with predefined topic labels and discovered distributions among these neurons. To be specific, Dai et al. (2021); Meng et al. (2022) identified knowledge neurons responsible for expressing factual knowledge. Wang et al. (2022b); Panigrahi et al. (2023) also discovered task-specific skill neurons, which are highly predictive of task performance and whose perturbation can drastically impair the corresponding ability. Gurnee et al. (2023); Voita, Ferrando, and Nalmpantis (2023) demonstrated certain neurons encode features like linguistics and positions. For multilingual processing, language-specific neurons have also been identified (Zhao et al. 2023a; Tang et al. 2024; Zhao et al. 2024b). By providing inputs designed to elicit varied functionalities, Zhang et al. (2023); Xiao et al. (2024) observed that specific neurons are responsible for distinct functions. Using SAEs to learn dictionary features for LLM activations, Templeton et al. (2024) found that the embeddings of the dictionary features form clusters that align with their semantics. Previous research has identified functional neuron sets, but these sets exhibited considerable overlap (Xiao et al. 2024), meaning truly decoupled modules remained undiscovered. In this paper, we discover and interpret the basic function modules, providing a clearer insight into the model’s internal functional organization.

3 Methodology

To discover function modules within LLMs, we first propose an optimization problem called *Neuron-Sample Dual Partitioning*. Subsequently, we propose an iterative partitioning algorithm to solve this problem.

3.1 Neuron-Sample Dual Partitioning Problem

An LLM consists of massive hidden units that can generate intermediate scalar values for various inputs, typically referred to as neurons. Our goal is to disentangle these neurons into several function modules, each characterized by dense co-activation on a specific group of samples that share a common topic.

Specifically, we assume an LLM consists of several latent function modules associated with a set of neurons and a set of representative samples that utilize it. However, what functions form decoupled modules in an LLM is initially unknown, making it infeasible to identify different function modules with neurons activated by manually selected samples. Consequently, we need to simultaneously discover both the neuron composition and the functional semantics of these decoupled modules. Formally, given a set of neurons $U = \{u_1, u_2, \dots, u_N\}$ in

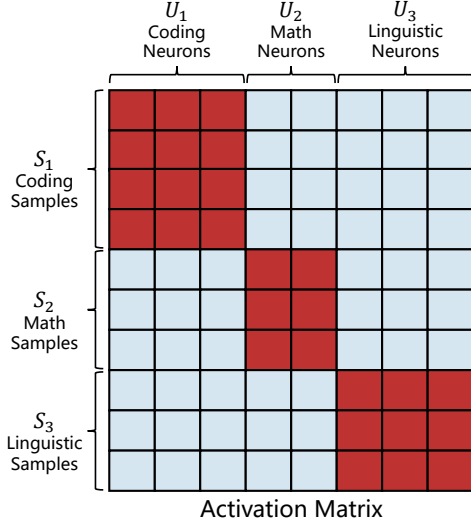


Figure 2: Illustration of function modules. Each red block represents a function module, associated with a set of neurons and a set of representative samples employing it.

an LLM. A set of input samples $S = \{s_1, s_2, \dots, s_M\}$ goes through the LLM and forms an activation matrix $A \in \mathbb{R}^{N \times M}$, where $A_{n,m}$ represents the activation value of neuron u_n for sample s_m . We aim to find K function modules $F = \{(S_1, U_1), (S_2, U_2), \dots, (S_K, U_K)\}$, where $\mathcal{P}_S = \{S_1, S_2, \dots, S_K\}$ and $\mathcal{P}_U = \{U_1, U_2, \dots, U_K\}$ are the samples and neurons that belong to each functional module, which is shown in Fig. 2

To simplify the problem, we restrict each sample and neuron assigned to one and only one functional module, representing the primary function they involve. This approach helps identify the overall semantics of functions through their core samples and neurons. With this semantic foundation established, we can then use activation analysis to discover additional function modules that each sample and neuron participates in.

Therefore, we formalize a Neuron-Sample Dual Partitioning optimization problem, *i.e.*, maximizing $\mathcal{L}(F)$ subject to:

- (I) **Completeness:** $\bigcup_{k=1}^K S_k = S, \quad \bigcup_{k=1}^K U_k = U,$
- (II) **Exclusivity:** $\forall k \neq j, \quad S_k \cap S_j = \emptyset, \quad U_k \cap U_j = \emptyset,$
- (III) **Non-Emptiness:** $\forall k \in [1, K], \quad S_k \neq \emptyset, \quad U_k \neq \emptyset,$

where $\mathcal{L}(F)$ is the objective function optimizing the modularity of function modules F , which is described subsequently.

3.2 Optimization Objectives

Intuitively, an ideal partition should exhibit strong activation within each module and a relatively balanced distribution of module sizes. Therefore, $\mathcal{L}(F)$ combines two aspects: Activation Modularity $\xi(F)$ and Balance Score $B(F)$.

Activation Modularity $\xi(F)$ We hypothesize that neurons and samples belonging to the same functional module should

Algorithm 1: IterD for Function Modules Discovery

Input: Activation Matrix A , Sample Set S , Neuron Set U

Hyper-Parameters: Number of Modules K

Output: $F = \{(S_k, U_k)\}_{k=1}^K$ that maximizes $\mathcal{L}(F)$

- 1: Initialize partition F_0 .
- 2: $t \leftarrow 0$.
- 3: **repeat**
- 4: Let $\{(S_k^t, U_k^t)\}_{k=1}^K = F_t$.
- 5: **{Step 1: Optimize Neuron Assignments}**
- 6: Create a temporary neuron partition $\mathcal{P}'_U = \{U'_1, \dots, U'_K\} = \{U_k^t\}_{k=1}^K$.
- 7: **for** each neuron $u \in U$ **do**
- 8: Compute $k_u^* \leftarrow \arg \max_{k \in \{1, \dots, K\}} \mathcal{L}(F_{u \rightarrow k})$ based on S_k^t .
- 9: Assign u to its new group in \mathcal{P}'_U .
- 10: **end for**
- 11: **{Step 2: Optimize Sample Assignments}**
- 12: Create a temporary sample partition $\mathcal{P}'_S = \{S'_1, \dots, S'_K\} = \{S_k^t\}_{k=1}^K$.
- 13: **for** each sample $s \in S$ **do**
- 14: Compute $k_s^* \leftarrow \arg \max_{k \in \{1, \dots, K\}} \mathcal{L}(F_{s \rightarrow k})$ based on U'_k .
- 15: Assign s to its new group in \mathcal{P}'_S .
- 16: **end for**
- 17: Let $F_{t+1} \leftarrow \{(S'_k, U'_k)\}_{k=1}^K$.
- 18: **until** F_t is identical to F_{t-1}
- 19: **return** F_t

exhibit an overall higher activation magnitude. Intuitively, a high activation magnitude implies that a neuron has a significant impact on processing a particular sample. We therefore define an objective to maximize the average activation within modules, which we term *Activation Modularity* $\xi(F)$.

Since computation in FFNs is token-wise, we first denote the activation of neuron u_i for sample s_j at token t as $a_{i,j,t}$. The average activation magnitude of neuron u_i on sample s_j is then defined as the mean of the absolute normalized activations across all tokens: $A_{i,j} = \frac{1}{T_j} \sum_{t=1}^{T_j} |a_{i,j,t}|$, where T_j is the number of tokens in sample s_j . Besides, to ensure comparable activation magnitudes across neurons, each neuron's activation is normalized by *z-score* over all samples.

For a modules' partition F containing K function modules (S_k, U_k) , $\xi(F)$ is the mean activation across all neuron-sample pairs (u, s) that are grouped in the same module:

$$\xi(F) = \frac{\sum_{k=1}^K \sum_{u_i \in U_k, s_j \in S_k} A_{i,j}}{\sum_{k=1}^K |U_k| |S_k|}. \quad (1)$$

A higher $\xi(F)$ indicates stronger internal activation within the modules, suggesting that it has successfully grouped neurons and samples with strong functional relations.

Balance Score $B(F)$ In practice, directly optimizing for $\xi(F)$ can lead to severely imbalanced module sizes. For example, it might result in one very large module and many very small ones, even if such a neuron partitioning achieves a high $\xi(F)$. To promote balanced module sizes and penalize

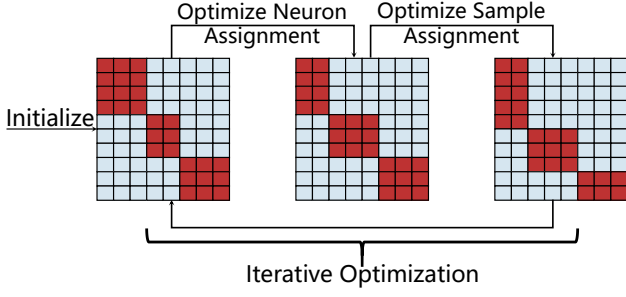


Figure 3: Overview of the IterD optimization process.

fragmentation, we define the balance score $B(F)$:

$$B(F) = \frac{K}{\sum_{k=1}^K \frac{1}{|U_k||S_k|}}. \quad (2)$$

A high $B(F)$ means the module sizes are balanced. We employ the harmonic mean here because it is particularly sensitive to small values. Consequently, this balance score strongly penalizes partitions where one or more modules become trivially small, ensuring that all K modules are meaningful. Finally, our objective function $\mathcal{L}(F)$ is defined as the product of the Activation Modularity and the Module Balance Coefficient: $\mathcal{L}(F) = \xi(F) \cdot B(F)$. Our fundamental goal is to identify a high-quality partition F that maximizes $\mathcal{L}(F)$, achieving both *strong internal activation* and a *balanced distribution of module sizes*.

3.3 Functional Region Identification Algorithm

Directly maximizing $\mathcal{L}(F)$ is computationally challenging due to the combinatorial nature of the partitioning problem. We, therefore, propose IterD, an iterative algorithm designed to find a high-quality partition F . The process, illustrated in Fig. 3, consists of an initialization stage (**Stage I**) followed by an iterative optimization stage (**Stage II**).

In **Stage I**, IterD begins with an initial partition $F_0 = \{(S_k^0, U_k^0)\}_{k=1}^K$. This can be generated by applying a basic clustering method (e.g., K-Means) to the samples and neurons to establish a reasonable starting point.

Stage II is iterative optimization. Given a partition $F_t = \{(S_k^t, U_k^t)\}_{k=1}^K$ at iteration t (with F_0 from initialization), IterD iteratively refines sample and neuron assignments to greedily maximize $\mathcal{L}(F)$. Each iteration involves two steps, which will be detailed below part by part:

Step 1: Optimize Neuron Assignments In this step, the sample partition $\mathcal{P}_S^t = \{S_k^t\}_{k=1}^K$ is held fixed. IterD iterates through each neuron $u_j \in U$. For each neuron, we determine its optimal new assignment by finding the target module k_j^* that greedily maximizes the objective function:

$$k_j^* = \arg \max_{k \in \{1, \dots, K\}} \mathcal{L}(F_{j \rightarrow k}), \quad (3)$$

where $F_{j \rightarrow k}$ is the resulting partition after reassigning neuron u_j to module k . The neuron is immediately reassigned to this new module before the next neuron is considered. Therefore, the optimization for each neuron is influenced by

the reassignments of all preceding neurons within the same step. After iterating through all neurons, this step yields an updated neuron partition $\mathcal{P}_U^{t+1} = \{U_k^{t+1}\}_{k=1}^K$.

Step 2: Optimize Sample Assignments Next, holding the newly updated neuron partition \mathcal{P}_U^{t+1} fixed, we perform a symmetric optimization for the samples. IterD iterates through each sample $s_i \in S$. For each sample, we find the target module k_i^* that greedily maximizes the objective $\mathcal{L}(F)$:

$$k_i^* = \arg \max_{k \in \{1, \dots, K\}} \mathcal{L}(F_{i \rightarrow k}), \quad (4)$$

where $F_{i \rightarrow k}$ denotes the hypothetical partition resulting from moving sample s_i to module k . As with the neurons, the sample is immediately reassigned. This process yields the updated sample partition $\mathcal{P}_S^{t+1} = \{S_k^{t+1}\}_{k=1}^K$. Putting these two steps together completes one full iteration, producing the new overall partition $F_{t+1} = \{(S_k^{t+1}, U_k^{t+1})\}_{k=1}^K$.

These two steps are repeated until convergence, i.e., when no neuron or sample reassignment occurs in a full iteration. More details are provided in Algo. 1 with a line-by-line explanation in Sec. ?? of Supp.

4 Experiments

In this section, we conducted extensive experiments to validate the effectiveness of our IterD framework (Sec. 3.3) and empirically provide some insights.

We used three open-sourced LLMs for fair comparisons: Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, and Qwen2.5-7B-Instruct. All-layer activations of these models were extracted from inference of some samples sourced from the Infinity-Instruct dataset (Li et al. 2025). These samples are balancedly selected following Xiao et al. (2024). More experimental details are provided in Sec. ?? of Supp.

We compare our framework against widely used clustering methods adapted for our task: **K-Means**, **Mini-Batch K-Means** (Sculley 2010), **Agglomerative Clustering** (Ward Jr 1963), **Spectral Clustering** (Ng, Jordan, and Weiss 2001), and **Spectral Co-clustering** (Dhillon 2001). We cluster the neurons with their activation vectors after PCA, i.e., row vectors $\text{PCA}(\{A_{i,:}\}_{i \in [1, N]})$. For some baselines with high complexities, we apply them on 10000 *sub-cluster centroids* pre-processed by K-Means.

4.1 Performance Comparison

We evaluate the performance of our framework and baselines via our optimization objective $\mathcal{L}(F)$ in Sec. 3.1 and a proposed novel Function Module Informativeness metric $\mathcal{I}(F)$, which comprehensively assesses the quality of the discovered function modules.

Comparison of $\mathcal{L}(F)$ As shown in Tab. 1, IterD consistently and substantially outperforms all baselines in optimizing $\mathcal{L}(F)$ across all LLMs and cluster counts (K), stemming from both higher activation modularity ($\xi(F)$) and a more balanced size distribution ($B(F)$). Therefore, the function modules discovered by our framework strictly match the motivation of our task.

Recall in Sec. 3.1, we find K function modules $F = \{(S_1, U_1), (S_2, U_2), \dots, (S_K, U_K)\}$, where sample group

LLM	K	K-Means			Mini-Batch K-Means			Agglomerative			Spectral			Spectral Co-cluster			IterD (Ours)		
		$\mathcal{L}(F)$	$\xi(F)$	$B(F)$	$\mathcal{L}(F)$	$\xi(F)$	$B(F)$	$\mathcal{L}(F)$	$\xi(F)$	$B(F)$	$\mathcal{L}(F)$	$\xi(F)$	$B(F)$	$\mathcal{L}(F)$	$\xi(F)$	$B(F)$	$\mathcal{L}(F)$	$\xi(F)$	$B(F)$
Qwen2.5-1.5B-Instruct	5	19.2	0.409	46.9	23.4	0.406	57.6	10.1	0.311	32.4	9.6	0.335	28.7	24.0	0.380	63.2	31.4	0.465	67.5
	10	5.7	0.644	8.8	4.2	0.630	6.7	5.2	0.599	8.7	2.8	0.566	5.0	7.9	0.522	15.1	12.0	0.711	16.9
	15	0.9	0.829	1.1	1.9	0.752	2.6	1.9	0.655	2.9	0.7	0.750	0.9	3.8	0.583	6.4	6.6	0.892	7.5
	20	0.4	0.943	0.4	0.5	0.855	0.6	1.0	0.731	1.3	0.3	0.914	0.3	2.0	0.636	3.1	4.4	1.048	4.2
Qwen2.5-3B-Instruct	5	28.0	0.466	60.1	26.9	0.490	54.9	16.1	0.439	36.6	0.0	0.001	7.7	34.6	0.404	85.6	56.9	0.533	106.8
	10	12.9	0.747	17.2	7.1	0.631	11.3	7.5	0.624	12.1	0.6	0.219	2.6	12.4	0.570	21.7	20.8	0.784	26.5
	15	1.1	0.853	1.3	1.4	0.818	1.7	2.1	0.836	2.5	0.4	0.423	0.8	1.1	0.559	1.9	11.3	0.964	11.8
	20	0.6	0.974	0.7	0.7	0.931	0.7	0.3	0.940	0.3	0.3	0.564	0.5	0.1	0.659	0.1	7.3	1.107	6.6
Qwen2.5-7B-Instruct	5	29.6	0.400	74.1	32.8	0.418	78.3	15.0	0.311	48.3	1.0	0.137	7.6	43.9	0.362	121.1	64.6	0.445	145.2
	10	13.1	0.639	20.6	12.5	0.624	20.1	9.8	0.518	18.8	1.3	0.384	3.4	16.9	0.522	32.3	25.2	0.702	35.9
	15	2.0	0.797	2.6	2.0	0.777	2.6	0.3	0.743	0.4	1.0	0.556	1.7	7.7	0.566	13.6	13.8	0.870	15.9
	20	0.4	0.903	0.5	0.6	0.905	0.7	0.5	0.753	0.7	0.0	0.834	0.0	4.3	0.634	6.8	9.0	1.008	8.9

Table 1: Comparison of Clustering Metrics for different methods, LLMs, and number of clusters (K). Values for $\mathcal{L}(F)$ and $B(F)$ are scaled by $\times 10^6$.

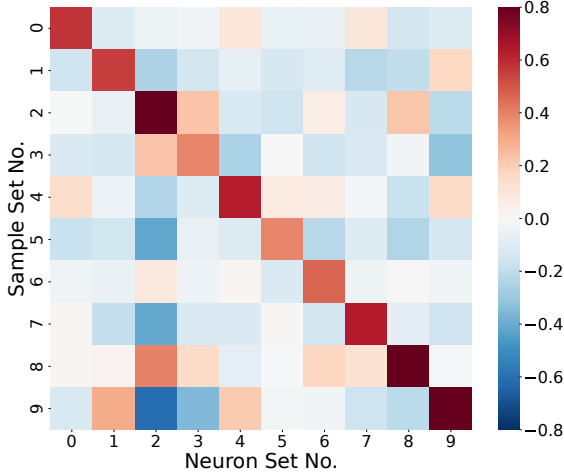


Figure 4: Average activation heatmap for discovered modules in Qwen2.5-7B-Instruct ($K = 10$). Each cell (i, j) shows the average activation of neuron set U_j on sample set S_i .

S_i and neuron group U_i should have a clear function correspondence. Here, we visualize such function correspondences in Fig. 4 for our modules discovered in Qwen2.5-7B-Instruct ($K = 10$), which can be well illustrated via the *co-activation patterns*, i.e., a highlighted structure in a co-activation heatmap. From Fig. 4, we can observe a clear block-diagonal highlighted structure (red), showing the discovered S_i and U_i are aligned functionally. In contrast, the non-diagonal low activations (blue/white) indicate minimal interactions between non-corresponding groups S_i and U_j ($i \neq j$). The strong comparison provides clear visual evidence that our algorithm successfully discovered *internally co-activating* and *functionally disentangled* modules in a fully unsupervised manner.

Comparison of $\mathcal{I}(F)$ To quantify the relevance between function modules and function labels, we propose a new metric called **Function Modular Informativeness**, denoted as $\mathcal{I}(F)$. For each sample $s_j \in S$, we construct an *activation pattern vector* $\mathbf{x}_{s_j} \in \mathbb{R}^K$, where its k -th component $\mathbf{x}_{s_j,k}$ de-

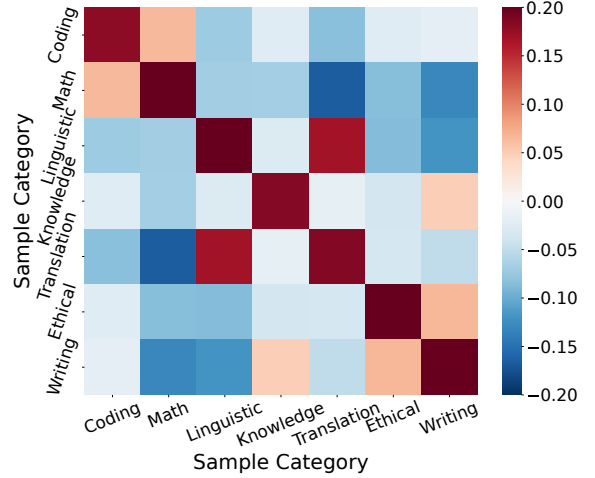


Figure 5: Visualization of sample category similarity in Qwen2.5-7B-Instruct ($K = 10$). The value in cell (i, j) represents the average cosine similarity between the feature vectors (\mathbf{x}_s) of samples from category i and category j .

notes how strongly s_j activates the k -th function modules U_k , i.e., $\mathbf{x}_{s_j,k} = \frac{1}{|U_k|} \sum_{u_i \in U_k} A_{i,j}$. It quantitatively describes s_j 's interactions with all identified modules. Specifically, we measure $\mathcal{I}(F)$ via the generalization performance of a linear classifier f_θ trained to predict the function labels y_j based on \mathbf{x}_{s_j} . We use either Logistic Regression (Cox 1958) or a linear Support Vector Classifier (SVC) (Cortes and Vapnik 1995). Formally,

$$\mathcal{I}(F) = \text{Acc}(f_{\theta^*}), \quad \theta^* = \arg \min_{\theta} \frac{1}{|S|} \sum_{s_j \in S} l(f_\theta(\mathbf{x}_{s_j}), y_j), \quad (5)$$

where $l(\cdot)$ is the standard classification loss, and $\text{Acc}(\cdot)$ denotes the accuracy or F1 score on the unseen test samples.

As shown in Tab. 2, our induced feature vectors \mathbf{x}_s achieve consistently higher classification accuracies and macro-F1 scores among all model sizes and values of K , regardless of whether a downstream Logistic Regressor or Support Vector classifier f_θ . Since such feature vectors are derived from

LLM	Cls.	K	K-Means		Mini-Batch K-Means		Agglomerative		Spectral		Spectral Co-Clustering		IterD (Ours)	
			Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5 -1.5B -Instruct	LR	5	0.4800	0.4758	0.5400	0.5404	0.5412	0.5409	0.5379	0.5300	0.4650	0.4602	0.5436	0.5428
		10	0.7186	0.7173	0.6929	0.6927	0.7057	0.7039	0.7150	0.7150	0.6793	0.6764	0.7257	0.7259
		15	0.7664	0.7662	0.7729	0.7724	0.7750	0.7750	0.7486	0.7474	0.7107	0.7089	0.8157	0.8158
		20	0.7907	0.7908	0.7843	0.7839	0.7850	0.7849	0.7864	0.7862	0.7643	0.7637	0.8200	0.8200
	SVC	5	0.5414	0.5417	0.5800	0.5829	0.5921	0.5933	0.5764	0.5695	0.5343	0.5339	0.5964	0.5987
		10	0.7307	0.7290	0.6986	0.6987	0.7479	0.7465	0.7264	0.7260	0.6929	0.6909	0.7536	0.7529
		15	0.7771	0.7770	0.7793	0.7792	0.7843	0.7843	0.7643	0.7638	0.6950	0.6950	0.8179	0.8181
		20	0.7986	0.7984	0.7886	0.7881	0.7843	0.7844	0.7943	0.7943	0.7671	0.7666	0.8236	0.8238
Qwen2.5 -3B -Instruct	LR	5	0.4700	0.4663	0.5071	0.5023	0.4864	0.4777	0.3636	0.3599	0.5007	0.4959	0.5550	0.5533
		10	0.6907	0.6872	0.6964	0.6965	0.7264	0.7260	0.4914	0.4903	0.6436	0.6427	0.7429	0.7424
		15	0.7421	0.7407	0.7750	0.7751	0.7664	0.7660	0.5779	0.5787	0.7093	0.7093	0.8171	0.8171
		20	0.7879	0.7876	0.7879	0.7875	0.7886	0.7889	0.7064	0.7065	0.7621	0.7620	0.8143	0.8143
	SVC	5	0.5336	0.5354	0.5500	0.5503	0.5414	0.5442	0.4150	0.4120	0.5593	0.5598	0.6014	0.6043
		10	0.7236	0.7207	0.7193	0.7199	0.7229	0.7225	0.5464	0.5514	0.6621	0.6623	0.7529	0.7519
		15	0.7479	0.7467	0.7743	0.7745	0.7593	0.7586	0.6029	0.6058	0.6957	0.6959	0.8150	0.8155
		20	0.7814	0.7815	0.7829	0.7828	0.7779	0.7779	0.7236	0.7229	0.7536	0.7531	0.8150	0.8151
Qwen2.5 -7B -Instruct	LR	5	0.5121	0.5086	0.5250	0.5254	0.5250	0.5194	0.3536	0.3483	0.4914	0.4886	0.6000	0.5992
		10	0.7071	0.7049	0.7336	0.7317	0.7300	0.7291	0.6136	0.6099	0.6486	0.6462	0.7536	0.7527
		15	0.7907	0.7900	0.7836	0.7827	0.7650	0.7634	0.7129	0.7116	0.7493	0.7484	0.8236	0.8236
		20	0.7929	0.7924	0.7986	0.7985	0.7893	0.7886	0.7729	0.7725	0.7779	0.7774	0.8179	0.8175
	SVC	5	0.5600	0.5610	0.5693	0.5709	0.5707	0.5706	0.4107	0.4077	0.5429	0.5436	0.6293	0.6297
		10	0.7357	0.7335	0.7586	0.7570	0.7564	0.7552	0.6536	0.6506	0.6650	0.6648	0.7750	0.7743
		15	0.7907	0.7901	0.7750	0.7742	0.7893	0.7884	0.7364	0.7353	0.7429	0.7419	0.8207	0.8206
		20	0.8000	0.7996	0.8036	0.8033	0.7993	0.7987	0.7607	0.7593	0.7621	0.7620	0.8214	0.8210

Table 2: Comparison of Test Accuracy and Macro-F1 for different methods, LLMs, classifiers, and number of clusters (K). Features are derived from average function module activations.

our discovered function modules, the superior downstream performance empirically proves that our algorithm discovers informative partitions with more meaningful functions.

Besides, to further investigate whether our neuron partition can structurally resemble ground-truth sample categories, motivated by contrastive learning, we analyzed *inter-category semantic disentanglement* $SD_{c,c'}$ and *intra-category semantic coherence* $SC_c = SD_{c,c'}$ via Cos-similarities between those activation pattern vectors (\mathbf{x}_s) with:

$$SD_{c,c'} = \frac{1}{|S^{[c]}| \cdot |S^{[c']}|} \sum_{s \in S^{[c]}} \sum_{s' \in S^{[c']}} \text{Cos}(x_s, x_{s'}), \quad (6)$$

where $S^{[c]}$ denotes the set of all samples of category $c \in [1, C]$. These $\{SD_{c,c'}\}$ values are visualized in Fig. 5, which reveals two findings: (1) Same-category samples (e.g., *Math* and *Coding*) exhibit highly similar activation patterns, whereas distinct-category samples display contrasting phenomena. (2) Some non-ignorable inter-category semantic connections can be successfully identified via those (corresponding) relatively high non-diagonal co-activations. For example, the notable semantic similarity $SD_{c,c'}$ between categories *Linguistic* and *Translation* achieves supervising 0.168, indicating reasonably similar behaviors of LLMs when dealing with such highly related tasks.

4.2 Empirical Insights of Function Modules

To provide qualitative insights into the nature of our identified function modules, we visualized the neuron groups for Qwen2.5-3B-Instruct with varying module numbers (i.e.,

$K \in \{10, 15, 20\}$) in Fig. 7, where each point represents a neuron and colors denote different function modules.

To further functionally analyze these modules, we first determine their specific functions from the LLMs’ summaries for some highly activated samples. We provide the details in Sec. ?? of Supp. For example, we inferred the function corresponding to the brown module should be *Algorithmic Programming*. The respective functions of other modules are labeled in Fig. 7. Besides, to study how LLMs perform or realize a function layer by layer, we illustrate in Fig. 6 how the neuron count of a function module is distributed in different layers.

Based on these figures, we observe several key findings:

Comprehensive Function Discovery All discovered modules correspond to interpretable functions tightly aligned with human knowledge or skill. Moreover, some critical functions consistently emerge across all levels of granularity K (such as Programming, Mathematics, and Writing), demonstrating the merits of our module partition algorithm.

Function Hierarchy Comparing the function visualizations for different values of K provides a clear hierarchical organization of such functions. A small K presents coarse-grained function disentanglement, while a large K gives a fine-grained function discovery akin to the “zoom in” effects. The examples described below evidently verify this insight:

- **Programming:** A large module *Algorithmic Programming* at $K = 10$ is split into several more specialized modules at $K = 20$ (e.g., *Code Analysis & Debugging*,

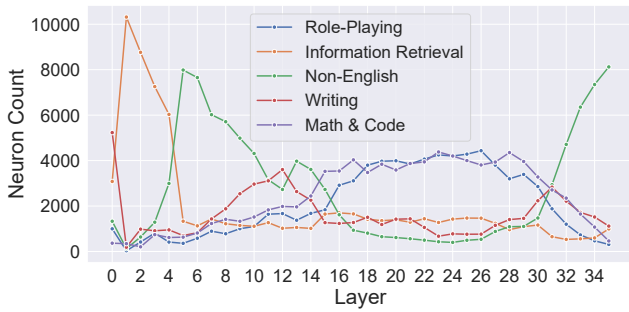


Figure 6: The studies of how neuron count is distributed in different layers inside each function module for Qwen2.5-3B-Instruct ($K = 5$).

and *Software & System Programming*), showing a clear divergence of engineering pathways.

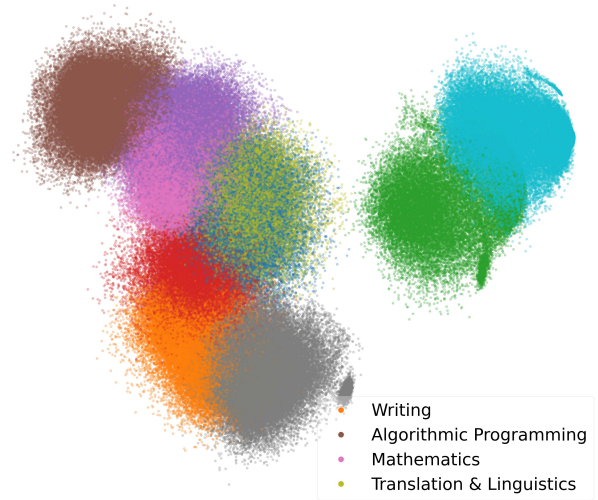
- **Writing:** The general *Writing* module from $K = 10$ is refined into *Creative Writing*, *Formal Writing*, and *Professional Writing* modules at $K = 20$, demonstrating an essential neuron hierarchy inside LLMs naturally supporting different writing styles.

Insights of Function Spatial Arrangement The locality of these modules is well aligned with their corresponding semantical functions. In other words, semantically related functions are located closely, revealing an intrinsic *connected* structure of related skills. For instance, functions related to programming, math, and science always appear nearby, forming a large "Techniques" connected component. In contrast, distant clusters often show low semantic relevance. Furthermore, a central hub of complex cognitive tasks (e.g., *Linguistics* and *Translation*) often emerges at the intersection of other modules, suggesting that they are core capabilities leveraged across multiple domains.

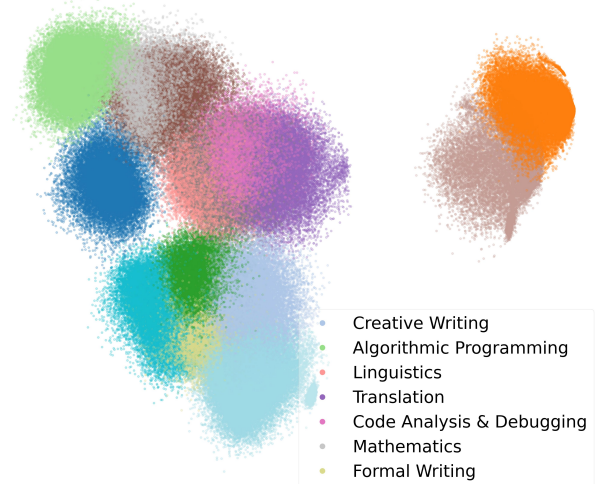
Layer Distribution of Functions From Fig. 6, different layers contribute quite variably to a function, which depends on the function's cognitive complexity. For instance, *Information Retrieval* requires only lower-layer processing, whereas functions like *Math & Code* require deeper-layer analysis.

5 Conclusion

In this paper, we propose a novel framework for discovering decoupled function modules in Large Language Models, addressing the challenge that prior work often identifies functionally overlapping neuron sets. We propose IterD, an iterative optimization algorithm, which maximizes a new objective function that prioritizes *high internally co-activations* and *clear functionally disentanglement* simultaneously in a fully unsupervised manner. On the Qwen2.5 LLMs family, our IterD framework significantly outperforms standard baselines. Moreover, our qualitative analysis reveals that the discovered modules show function comprehensiveness, function hierarchy, and clear function spatial arrangement within LLMs. Our work provides a novel tool for interpreting LLMs' function modules, filling a critical gap in LLMs' interpretability research.



(a) Function modules for $K = 10$



(b) Function modules for $K = 15$



(c) Function modules for $K = 20$

Figure 7: Visualization of discovered function modules.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (No. 62306255, 92370204), the National Key Research and Development Program of China (No. 2023YFF0725000), the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011839), and the Education Bureau of Guangzhou Municipality.

References

- Ameisen, E.; Lindsey, J.; Pearce, A.; Gurnee, W.; Turner, N. L.; Chen, B.; Citro, C.; Abrahams, D.; Carter, S.; Hosmer, B.; Marcus, J.; Sklar, M.; Templeton, A.; Bricken, T.; McDougall, C.; Cunningham, H.; Henighan, T.; Jermyn, A.; Jones, A.; Persic, A.; Qi, Z.; Ben Thompson, T.; Zimmerman, S.; Rivoire, K.; Conerly, T.; Olah, C.; and Batson, J. 2025. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread*.
- Bills, S.; Cammarata, N.; Mossing, D.; Tillman, H.; Gao, L.; Goh, G.; Sutskever, I.; Leike, J.; Wu, J.; and Saunders, W. 2023. Language models can explain neurons in language models. <https://openai-public.blob.core.windows.net/neuron-explainer/paper/index.html>.
- Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; Lasenby, R.; Wu, Y.; Kravec, S.; Schiefer, N.; Maxwell, T.; Joseph, N.; Hatfield-Dodds, Z.; Tamkin, A.; Nguyen, K.; McLean, B.; Burke, J. E.; Hume, T.; Carter, S.; Henighan, T.; and Olah, C. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Bullmore, E.; and Sporns, O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3): 186–198.
- Conmy, A.; Mavor-Parker, A.; Lynch, A.; Heimersheim, S.; and Garriga-Alonso, A. 2023a. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36: 16318–16352.
- Conmy, A.; Mavor-Parker, A. N.; Lynch, A.; Heimersheim, S.; and Garriga-Alonso, A. 2023b. Towards automated circuit discovery for mechanistic interpretability. URL <https://arxiv.org/abs/2304.14997>, 2.
- Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20: 273–297.
- Cox, D. R. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2): 215–232.
- Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.
- Dhillon, I. S. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 269–274.
- Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; DasSarma, N.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2021/framework/index.html>.
- Geva, M.; Bastings, J.; Filippova, K.; and Globerson, A. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.
- Gong, Z.; and Sun, Y. 2024. Graph reasoning enhanced language models for text-to-sql. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2447–2451.
- Guo, Y.; Guo, S.; Zhu, H.; and Sun, Y. 2025. Towards Lifelong Model Editing via Simulating Ideal Editor. In *Proceedings of the 42th International Conference on Machine Learning*.
- Gurnee, W.; Nanda, N.; Pauly, M.; Harvey, K.; Troitskii, D.; and Bertsimas, D. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.
- Hanna, M.; Liu, O.; and Variengien, A. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36: 76033–76060.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.
- Ji, Y.; Sun, Y.; Zhang, Y.; Wang, Z.; Zhuang, Y.; Gong, Z.; Shen, D.; Qin, C.; Zhu, H.; and Xiong, H. 2025. A comprehensive survey on self-interpretable neural networks. *arXiv preprint arXiv:2501.15638*.
- Kandel, E. R.; Schwartz, J. H.; Jessell, T.; Siegelbaum, S. A.; and Hudspeth, A. 2013. Principles of neural science.
- Li, J.; Du, L.; Zhao, H.; Zhang, B.-w.; Wang, L.; Gao, B.; Liu, G.; and Lin, Y. 2025. Infinity Instruct: Scaling Instruction Selection and Synthesis to Enhance Language Models. *arXiv preprint arXiv:2506.11116*.
- Lindsey, J.; Gurnee, W.; Ameisen, E.; Chen, B.; Pearce, A.; Turner, N. L.; Citro, C.; Abrahams, D.; Carter, S.; Hosmer, B.; Marcus, J.; Sklar, M.; Templeton, A.; Bricken, T.; McDougall, C.; Cunningham, H.; Henighan, T.; Jermyn, A.; Jones, A.; Persic, A.; Qi, Z.; Thompson, T. B.; Zimmerman, S.; Rivoire, K.; Conerly, T.; Olah, C.; and Batson, J. 2025. On the Biology of a Large Language Model. *Transformer Circuits Thread*.
- Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35: 17359–17372.
- Meunier, D.; Lambiotte, R.; and Bullmore, E. T. 2010. Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, 4: 200.
- Ng, A.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

- Olsson, C.; Elhage, N.; Nanda, N.; Joseph, N.; DasSarma, N.; Henighan, T.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Johnston, S.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2022. In-context Learning and Induction Heads. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- Panigrahi, A.; Saunshi, N.; Zhao, H.; and Arora, S. 2023. Task-specific skill localization in fine-tuned language models. In *International Conference on Machine Learning*, 27011–27033. PMLR.
- Sculley, D. 2010. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, 1177–1178.
- Stolfo, A.; Belinkov, Y.; and Sachan, M. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. *arXiv preprint arXiv:2305.15054*.
- Sun, Y.; Zhu, H.; Qin, C.; Zhuang, F.; He, Q.; and Xiong, H. 2021. Discerning decision-making process of deep neural networks with hierarchical voting transformation. *Advances in Neural Information Processing Systems*, 34: 17221–17234.
- Sun, Y.; Zhu, H.; and Xiong, H. 2025. Toward Faithful Neural Network Intrinsic Interpretation With Shapley Additive Self-Attribution. *IEEE Transactions on Neural Networks and Learning Systems*.
- Tang, T.; Luo, W.; Huang, H.; Zhang, D.; Wang, X.; Zhao, X.; Wei, F.; and Wen, J.-R. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.
- Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; Cunningham, H.; Turner, N. L.; McDougall, C.; MacDiarmid, M.; Freeman, C. D.; Sumers, T. R.; Rees, E.; Batson, J.; Jermyn, A.; Carter, S.; Olah, C.; and Henighan, T. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*.
- Voita, E.; Ferrando, J.; and Nalmpantis, C. 2023. Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827*.
- Wang, K.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2022a. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.
- Wang, S.; Zhu, Y.; Liu, H.; Zheng, Z.; Chen, C.; and Li, J. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3): 1–37.
- Wang, X.; Wen, K.; Zhang, Z.; Hou, L.; Liu, Z.; and Li, J. 2022b. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*.
- Ward Jr, J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301): 236–244.
- Xiao, C.; Zhang, Z.; Song, C.; Jiang, D.; Yao, F.; Han, X.; Wang, X.; Wang, S.; Huang, Y.; Lin, G.; et al. 2024. Configurable foundation models: Building llms from a modular perspective. *arXiv preprint arXiv:2409.02877*.
- Xin, H.; Sun, Y.; Wang, C.; and Xiong, H. 2025. Llmcdsr: Enhancing cross-domain sequential recommendation with large language models. *ACM Transactions on Information Systems*.
- Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
- Yao, Y.; Wang, P.; Tian, B.; Cheng, S.; Li, Z.; Deng, S.; Chen, H.; and Zhang, N. 2023. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*.
- Zhang, F.; and Nanda, N. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.
- Zhang, Z.; Zeng, Z.; Lin, Y.; Xiao, C.; Wang, X.; Han, X.; Liu, Z.; Xie, R.; Sun, M.; and Zhou, J. 2023. Emergent modularity in pre-trained transformers. *arXiv preprint arXiv:2305.18390*.
- Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024a. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38.
- Zhao, J.; Zhang, Z.; Ma, Y.; Zhang, Q.; Gui, T.; Gao, L.; and Huang, X. 2023a. Unveiling a core linguistic region in large language models. *arXiv preprint arXiv:2310.14928*.
- Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; et al. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).
- Zhao, Y.; Zhang, W.; Chen, G.; Kawaguchi, K.; and Bing, L. 2024b. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.