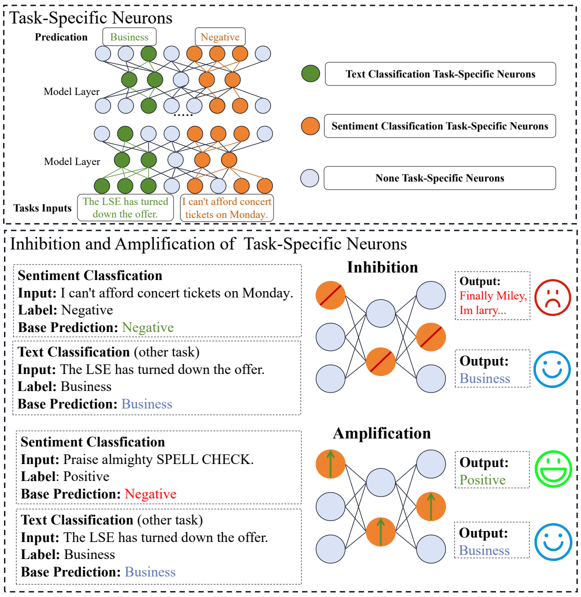
Does Large Language Model Contain Task-Specific Neurons?

Ran Song[1,2], **Shizhu He**[3,4], **Shuting Jiang**[1,2], **Yantuan Xian**[1,2],
Shengxiang Gao[1,2], **Kang Liu**[3,4], and **Zhengtao Yu**[1,2*],
[1] Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China
[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
[3] The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4] School of Artificial Intelligence, University of Chinese Academy of Science, Beijing, China
{song_ransr,shuting_jiang22}@163.com, {shizhu.he,kliu}@nlpr.ia.ac.cn,
xianyt@kust.edu.cn, {gaoshengxiang.yn,ztyu}@hotmail.com

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in comprehensively handling various types of natural language processing (NLP) tasks. However, there are significant differences in the knowledge and abilities required for different tasks. Therefore, it is important to understand whether the same LLM processes different tasks in the same way. Are there specific neurons in a LLM for different tasks? Inspired by neuroscience, this paper pioneers the exploration of whether distinct neurons are activated when a LLM handles different tasks. Compared with current research exploring the neurons of language and knowledge, task-specific neurons present a greater challenge due to their abstractness, diversity, and complexity. To address these challenges, this paper proposes a method for task-specific neuron localization based on Causal Gradient Variation with Special Tokens (CGVST). CGVST identifies task-specific neurons by concentrating on the most significant tokens during task processing, thereby eliminating redundant tokens and minimizing interference from non-essential neurons. Compared to traditional neuron localization methods, our approach can more effectively identify task-specific neurons. We conduct experiments across eight different public tasks. Experiments involving the inhibition and amplification of identified neurons demonstrate that our method can accurately locate task-specific neurons.


Task-Specific Neurons
Predication
Business
Negative
Model Layer
.....
Model Layer
Tasks Inputs
The LSE has turned down the offer.
I can't afford concert tickets on Monday.
Text Classification Task-Specific Neurons
Sentiment Classification Task-Specific Neurons
None Task-Specific Neurons
Inhibition and Amplification of Task-Specific Neurons
Inhibition
Amplification

Sentiment Classfication
Input: I can't afford concert tickets on Monday.
Label: Negative
Base Prediction: Negative

Output: Finally Miley, Im larry...

Text Classification (other task)
Input: The LSE has turned down the offer.
Label: Business
Base Prediction: Business

Output: Business

Sentiment Classfication
Input: Praise almighty SPELL CHECK.
Label: Positive
Base Prediction: Negative

Output: Positive

Text Classification (other task)
Input: The LSE has turned down the offer.
Label: Business
Base Prediction: Business

Output: Business

Figure 1: The upper shows different task-specific neurons from the same LLM. The bottom checks if these neurons affect the tasks by inhibiting and amplifying.

1 Introduction

Large Language Models (LLMs) have gained widespread attention due to their powerful capabilities (Zhao et al., 2023b). Based on the unsupervised pre-training followed by instruction fine-tuning (IFT) paradigm (Brown et al., 2020), LLMs have developed the ability to comprehensively handle various types of natural language processing (NLP) tasks. The advantage of this paradigm is that a single model deployment can perform multiple tasks, showcasing the versatility and efficiency of LLMs in NLP applications (Yuan et al., 2024).

However, there are significant differences in the knowledge and abilities required for different tasks by LLMs. For example, sentiment analysis focuses on adjectives and adverbs (Benamara et al., 2007), text classification emphasizes domain-specific terminology (Avancini et al., 2006), while natural language inference prioritizes the relationship between premise and conclusion sentences (Camburu et al., 2018). This suggests that the processing patterns of the same LLM may vary across different tasks, highlighting the necessity to investigate

[*] Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7101–7113
November 12-16, 2024 ©2024 Association for Computational Linguistics

the underlying mechanisms of task-specific processing in LLM. Task-specific processing can be traced by examining the neurons activated during the inference of specific tasks. Therefore, the two critical questions emerge: *Are there neurons within the same LLM that handle specific tasks?*, and if so, *how does the model manage and differentiate between various types of tasks?*

In fact, neuroscience has discovered that different brain areas control distinct behavioral abilities (Bari and Robbins, 2013). For instance, while higher cognitive functions such as learning, reasoning, decision-making and creativity are primarily controlled by the frontal lobe, the neurons involved in these processes remain distinct (Collins and Koechlin, 2012). Inspired by these findings, we hypothesize that although LLMs utilize a unified structure and parameters, the neurons engaged in different tasks may vary significantly. Motivated by this hypothesis, this paper investigates the existence of task-specific neurons in LLMs, specifically examining whether the neurons activated by different tasks exhibit distinct patterns. As shown in Figure 1 upper, different task-specific neurons in the same LLM are marked with different colors.

Current research has explored specific neurons in LLMs that primarily handle knowledge (Dai et al., 2022; Niu et al., 2024; Chen et al., 2024) and languages (Zhao et al., 2023a; Tang et al., 2024). On the one hand, knowledge-specific neurons in LLMs are key units that store factual information. By adjusting the activity of these neurons, we can control how much certain knowledge is represented in the model (Dai et al., 2022). This ability to manipulate knowledge-specific neurons enables knowledge editing, allowing for the modification of the factual information stored in the model (Meng et al., 2022a,b). On the other hand, language-specific neurons in LLMs are specialized units that focus on language-related tasks, such as language modeling and machine translation. These neurons are responsible for controlling the quality of language generation in the model (Zhao et al., 2023a). By manipulating the activity of language-specific neurons, it is possible to influence the way LLMs generate language, such as translations or responses in different languages. For example, researchers have shown that adjusting these neurons can enable LLMs to switch between languages more effectively (Tang et al., 2024). Utilizing language-specific neurons can help LLMs better handle linguistic tasks.

Moreover, current methods cannot directly explore task-specific neurons because they have the following main different characteristics from language- and knowledge-specific neurons: 1) **Abstractness**: Unlike specific knowledge and language, the competencies required for tasks are more intricate and challenging to represent within symbolic systems. 2) **Diversity**: While knowledge and language can be examined through a finite number of enumerated examples, task examples exhibit diversity and are difficult to comprehensively enumerate. 3) **Complexity**: The abilities necessary for tasks may exhibit interdependencies or operate independently, often lacking clear demarcations.

Specifically, we found that task-specific neurons only need to solve the target task; for example, sentiment analysis neurons do not need to contain general English neurons. Thus, we believe that not all tokens are crucial for identifying task-specific neurons. A similar approach has been used in tasks such as continuous learning (Lin et al., 2024) and computer vision (Zeng et al., 2022). Different parts of inputs, such as task definitions, and contextual examples, play distinct roles in the inference process over LLMs (Jiang et al., 2023b). In biology, studies often use fluorescent markers to track neuronal activity by injecting fluorescent agents and observing their effects to monitor neurons (Chen et al., 2013).

Inspired by this, we propose a novel task-specific neuron detection method using Causal Gradient Variation with Special Tokens (CGVST). We introduced fluorescent markers into the LLM reasoning process to uncover the crucial role of special tokens. By observing neuron activation patterns triggered by these tokens, we can obtain abstract semantic representations beyond individual examples. This approach also helps to control diversity and complexity by focusing only on the special tokens. Finally, we perform causal masked language model prediction for the task, recording the gradient on special tokens, and identifying the most active neurons as the task-specific neurons.

We also conduct detailed experiments on 8 different tasks (Wang et al., 2022b), including quantitative experiments on neuron inhibition and amplification. Quantitative experiments demonstrate that our method can accurately identify task-specific neurons. As shown in Figure 1 bottom, inhibition of these neurons significantly reduces performance on the target task, while the effect on other tasks is minimal. Conversely, amplifying these neurons improves performance in the related tasks with minimal effect on other tasks. Furthermore, we conducted comprehensive analyses of the neurons, including cross-validation and neuron visualization. The codes for this paper are available at GitHub[1].

Our contributions can be summarized concisely in the following three aspects:

- We propose the existence of task-specific neurons, defining them as neurons that have a significant impact on a specific task while having minimal impact on other tasks.
- We introduce a method to identify task-specific neurons based on Causal Gradient Variation with Special Tokens (CGVST). By analyzing the importance of special tokens and recording gradient variation at them during task processing, we identify the most relevant task-specific neurons.
- We conducted several analytical experiments on task-specific neurons. The results show that our method can effectively locate them, and they align more closely with the definition of task-specific neurons compared to other methods.

2 Background

2.1 Memory Mechanism of FFN

Transformer is an efficient network architecture employed in various tasks (Vaswani et al., 2017). Mainstream LLMs often utilize multi-layer Transformer decoders. Each Transformer layer comprises two components: Multi-Head Self-Attention (MHA) and Feed-Forward Network (FFN). Taking the mainstream LLMs LLama (Touvron et al., 2023) and Mistral (Jiang et al., 2023a) as examples, the Multi-Head Self-Attention mechanism can be expressed as follows:

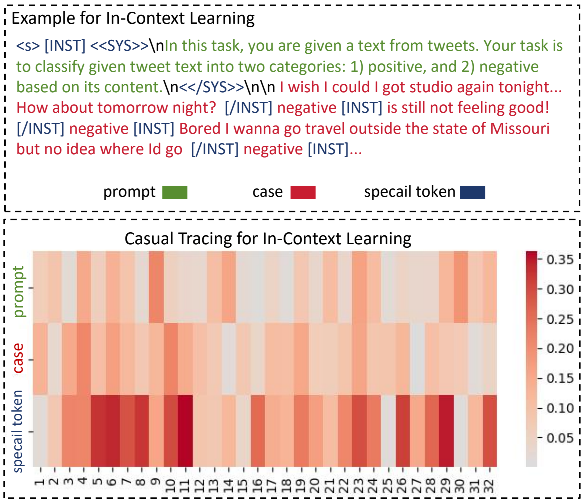
$$\mathbf{h}_a^l = \mathrm{Attn}(\mathbf{h}^{l-1}\mathbf{W}_q^l, \mathbf{h}^{l-1}\mathbf{W}_k^l, \mathbf{h}^{l-1}\mathbf{W}_v^l) \cdot \mathbf{W}_o^l, \tag{1}$$

where $\mathbf{h}^{l-1}$ represents the output from the previous layer, $\mathbf{W}_q^l$, $\mathbf{W}_k^l$, and $\mathbf{W}_v^l$ denote the weight matrices for queries, keys, and values, respectively, and $\mathbf{W}_o^l$ is the output weight matrix.

Next, Feed-Forward Network (FFN) is represented as follows:

$$\mathbf{h}^{l+1} = f_{act}(\mathbf{h}_a^l\mathbf{W}_{gate}^l) \circ \mathbf{h}_a^l\mathbf{W}_{up}^l \cdot \mathbf{W}_{down}^l, \tag{2}$$

where $\mathbf{W}_{up}^l$ and $\mathbf{W}_{down}^l$ are the weight matrices, $\mathbf{W}_{gate}^l$ represents the gating parameter, and f_{act} is the activation function, such as ReLU (Agarap, 2019) or SiLU (Elfwing et al., 2017). Studies have shown that language patterns and knowledge are memorized in the FFN layer, and these memories are triggered by modulating the activation state (Geva et al., 2021). Therefore, $\mathbf{W}_{gate}^l$ can be inferred that memory in LLMs is derived from this gating mechanism, which plays a crucial role in regulating activation states within the FFN layers.



Example for In-Context Learning

<s> [INST] <<SYS>>\nIn this task, you are given a text from tweets. Your task is to classify given tweet text into two categories: 1) positive, and 2) negative based on its content.\n<</SYS>>\n\n I wish I could I got studio again tonight... How about tomorrow night? [/INST] negative [INST] is still not feeling good! [/INST] negative [INST] Bored I wanna go travel outside the state of Missouri but no idea where Id go [/INST] negative [INST]...

Figure 2: The top part of the figure displays examples of ICL and the various roles involved. The bottom part illustrates the causal tracing results for these roles on the Y-axis. The X-axis represents the model's layers, with darker colors indicating a more significant impact on task performance.

2.2 In-Context Learning

With the advancing capabilities of LLMs, In-Context Learning (ICL) has emerged as a new paradigm in NLP tasks (Dong et al., 2022). In this approach, LLMs make predictions based solely on contexts augmented with a prompt, a few examples and special tokens, as shown in Figure 2. The process can be described as:

$$\hat{y} = \arg\max_{y_j \in Y} P(y_j \mid x_j, P, C, S) \tag{3}$$

where C represents the context and is defined as a set of pairs: $C \in \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. The prediction space is denoted by Y, while P is the prompt that defines the task. Special tokens play a crucial role in regulating LLMs behavior within the Chain of Thought framework. They help to manage multiple rounds of dialogue, enabling more coherent and contextually relevant responses. By using these tokens, LLMs can maintain context

[1] https://github.com/Maxpa1n/task-neurons

over longer content and complex interactions, enhancing performance in various NLP tasks.

3 Task-specific Neurons Identification

In this section, we propose a simple and innovative strategy to identify task-specific neurons. By tracking gradient changes of specific tokens, we can locate neurons essential for task processing.

3.1 Causal Tracing of Context

In inference with ICL, task prompts define tasks for LLMs, and contextual examples teach task processing. High-quality examples can enhance task performance (Zhang et al., 2023). Special tokens mark the boundaries between different roles of inputs and absorb their representations.

To evaluate the significance of these three roles, this study conducts a causal tracing analysis for each token. We distinguish the roles within an input $(X \in \{x_p, x_c, x_s\}, Y)$. During inference, we record the predicted probability p_Y of the correct label based on the LLMs' parameters θ,

$$p_Y = P_\theta(Y|x_p, x_c, x_s). \tag{4}$$

Subsequently, we add noise to each token at each FFN layer as a fluorescent vector and record the predicted probability for the correct label, to construct a perturbation matrix $\mathbf{A}$. The rows and columns of the matrix correspond to the indices of the FFN layers and the length of the inputs.

$$\mathbf{A}_{ij} = P_\theta(Y \mid x_p, x_c, x_s, \mathbf{h}_i^{*j}) \tag{5}$$

$$\mathbf{h}_i^{*j} = \mathbf{h}_i^j \mathbf{W}_{gate}^j + \epsilon \tag{6}$$

$$\mathbf{h}_i^{j+1} = f_{act}(\mathbf{h}_i^{*j}) \circ \mathbf{h}_i^j \mathbf{W}_{up}^j \cdot \mathbf{W}_{down}^j \tag{7}$$

where $\mathbf{A}_{ij}$ represents the noise-added prediction of the i-th token of the j-th layer, $\mathbf{h}_i^{*j}$ is the fluorescent vector, and $\epsilon \in \mathcal{N}(0, 1)$.

Next, we calculate the difference between the values in the perturbation matrix and the original confidence, determining the significance of each token in each layer for the task. We categorize these probabilities according to different roles to identify which types of tokens most influence the task. As shown in Figure 2, we find that perturbing special tokens has the most significant effect on task performance. Therefore, we suggest that the effectiveness of LLMs in task processing is mainly attributed to the representation of special tokens.

3.2 Causal Gradient Variation with Special Tokens

Following the analysis, we discover that the gating parameters in the FFN layer store the task pattern memory. This means the model's ability to handle different tasks is encoded in a specific memory structure, allowing for efficient task-switching and performance optimization. Using causal tracing, we identify that special tokens in the context are crucial for task performance. These token representations contain the patterns for task processing. Building on these insights, we propose a novel method to identify task-specific neurons based on Causal Gradient Variation with Special Tokens (CGVST). This method uses the gradient of special tokens to find neurons that are particularly sensitive to specific tasks.

Initially, we perform a forward pass using task-specific data, focusing on computing the loss function when special tokens are predicted, as follows:

$$\mathcal{L}_s^i = -\sum_{i=s}^{T_{x_s}} \log P(x_i \mid x_1, x_2, \ldots, x_{i-1}, \\ x_i \in x_s) \tag{8}$$

where s denotes the position of the special tokens, and T_{x_s} represents the special token set.

Next, we calculate the gradient variation on the training set for the specified task. The gradient of the gated weight is determined based on these loss values, as shown below:

$$\delta = \sum_{i}^{T_{x_s}} \frac{\partial \mathcal{L}_s^i}{\partial \mathbf{W}_{gate}} \tag{9}$$

where $\delta \in \mathbb{R}^{l \times d \times 4d}$, with l is number of layer, and d being the dimensionality of each layer.

Given that the size of the FFN gate value for each layer is $4d$-dimensional, we compressed the gradient changes to d-dimensions to obtain a matrix of the same size as $\mathbf{h}\mathbf{W}_{gate}$. Subsequently, we select the n positions with the largest variations globally as the task-specific neurons, which are considered crucial for the task. A single neuron is denoted as α_i^j, representing the i-th neuron of the j-th layer.

3.3 Task Control by Task-Specific Neurons

Following the above steps, we can identify task-specific neurons. These neurons exhibit significant roles during task processing, indicating their critical function. Using these identified task-specific

	QA		SA		QU		LTC		TC		CEC		EC		TM		AVG		ICP
	P↓	R↑	P↓	R↑	P↓	R↑	P↓	R↑	P↓	R↑	P↓	R↑	P↓	R↑	P↓	R↑	P↓	R↑	-
BASE	38.7	59.3	68.8	55.0	64.9	55.6	76.0	54.0	66.2	55.4	54.4	57.1	37.4	59.5	47.7	58.0	56.8	56.8	-
RANDOM	38.9	57.8	69.4	52.2	63.9	57.0	70.3	52.1	62.2	56.2	54.5	51.3	37.1	52.8	47.8	54.3	55.3	54.2	-1.1
PV (Zhao et al., 2023a)	37.8	**39.4**	65.2	**53.2**	61.8	51.1	73.2	**51.2**	66.2	**53.2**	53.2	**51.9**	27.8	**34.2**	45.2	**47.7**	53.8	**47.7**	-6.1
LAPE (Tang et al., 2024)	31.7	3.6	59.6	49.4	41.4	46.9	56.7	46.9	22.9	23.9	54.3	50.0	25.8	20.3	45.9	41.7	42.3	35.3	-7.0
GV (Dai et al., 2022)	12.9	17.1	47.9	47.8	57.7	**54.5**	52.8	46.6	22.0	30.7	39.5	50.4	17.9	25.6	38.4	45.4	36.1	39.8	3.70
CGVST (ours)	**3.4**	27.3	**3.3**	18.4	**17.2**	31.4	**16.4**	24.4	**7.4**	29.4	**35.7**	29.8	**2.2**	19.7	**27.8**	34.3	**13.8**	26.8	**14.0**

Table 1: This table presents the results obtained after inhibiting task-specific neurons in various tasks. P denotes the accuracy of the inhibited task, while R indicates the performance on other tasks when the current task is inhibited. The Inhibition Comprehensive Performance (ICP) is calculated as $(BASE - P) - |BASE - R|$, balancing both P and R. Underline indicates that tasks are almost ineffective. **Bold** represents the best result of the indicator.

neurons, we can further manipulate them to influence the overall performance of LLMs. Specifically, we can amplify (by increasing activation levels) or inhibit (by decreasing activation levels) the activation value of task-specific neurons. The operation is as follows:

$$\mathbf{h}^{l+1} = f_{act}(\mathbf{h}_a^l \mathbf{W}_{gate}^l) \circ \alpha_*^l \circ \mathbf{h}_a^l \mathbf{W}_{up}^l \cdot \mathbf{W}_{down}^l, \tag{10}$$

where α_*^l represents the selected neuron in the l-th layer. This neuron is set to a value less than 1 when inhibited, greater than 1 when amplified, and equal to 1 during normal inference.

4 Experiments

In this section, we conduct a series of detailed experiments to investigate task-specific neurons. We introduce 8 distinct tasks, each serving as an objective for identifying specific neurons. The experiments aim to answer the following research questions: **RQ1**: Can the proposed method locate neurons? For the located neurons, does inhibiting (§ 4.4) and amplifying them (§ 4.5) have corresponding effects? **RQ2**: Do task-specific neurons impact the model's language ability? (§ 4.6). **RQ3**: What is the relationship between neurons corresponding to different tasks? (§ 4.7) And how are they distributed in the model? (§ 4.8).

4.1 Dataset

We selected 8 tasks from the Super-Natural Instruction dataset (Wang et al., 2022c)[2] as follows: **Question Answering (QA)** generates answers to SQuAD 1.1 questions based on documents. **Sentiment Analysis (SA)** classifies the sentiment of an English tweet as positive or negative in social media. **Question Understanding (QU)** determines if a clarification for a query is correct by responding with *Yes* or *No* in dialogue. **Text Categorization (TC)** classifies the topic of an English news article into one of four classes in news. **Law Text Categorization (LTC)** classifies an English sentence as either overruling or non-overruling in law. **Cause Effect Classification (CEC)** decides if the second sentence logically results from the first one in commonsense reasoning. **Emotion Classification (EC)** classifies the emotion of a Twitter post into one of six classes: sadness, joy, love, anger, fear, or surprise, in social media. **Text Matching (TM)** classifies pairs of medical questions into two categories in medicine and healthcare. All datasets are evaluated using Exact Match accuracy.

4.2 Baseline

We compared the current neuron selection methods with several baseline approaches. Language Activation Probability Entropy (LAPE) identifies the most active neurons during inference by calculating the entropy of activation frequency and value, considering these as task-relevant neurons (Tang et al., 2024). Parameter Variation (PV) involves training on the corresponding task and then identifying neurons with the least parameter changes pre- and post-training, which are considered most relevant to the task (Zhao et al., 2023a). Gradient Variation (GV) determines task-relevant neurons by finding the parameters with the largest gradients across all tokens during task training (Dai et al., 2022). The RANDOM method randomly selects neurons from different layers and positions.

4.3 Implementation Details

For each task, we split the data into two parts: one half for training to identify task-specific neurons, and the other half for testing to evaluate model performance. We utilize the LLama2-7b-chat model with a 5-shot ICL for inference. The total number of neurons is calculated as 32 (the number of model

[2] The numbers in the dataset are: 075, 195, 227, 274, 379, 391, 512, 1645.

	QA		SA		QU		LTC		TC		CEC		EC		TM		AVG		ACP
	P↑	R↑	P↑	R↑	P↑	R↑	P↑	R↑	P↑	R↑	P↑	R↑	P↑	R↑	P↑	R↑	P↑	R↑	-
BASE	38.7	59.3	68.8	55.0	64.9	55.6	76.0	54.0	66.2	55.4	54.4	57.1	37.4	59.5	47.7	58.0	56.8	56.8	-
RANDOM	37.9	56.8	65.4	54.2	62.2	54.7	71.4	53.4	66.1	55.1	55.1	52.2	38.4	56.4	47.8	56.7	55.53	54.93	-3.1
PV (Zhao et al., 2023a)	27.7	35.3	43.7	65.2	56.0	60.6	67.7	**70.2**	36.7	**59.7**	45.7	53.6	38.7	37.7	32.4	43.8	40.1	53.3	-20.2
LAPE (Tang et al., 2024)	39.1	38.3	63.0	64.9	**64.2**	**65.6**	62.1	60.3	45.1	46.2	54.3	53.7	31.0	32.4	45.3	40.1	48.8	50.2	-14.6
GV (Dai et al., 2022)	30.5	29.5	46.0	**67.7**	56.4	54.7	47.8	55.7	56.7	47.6	56.7	49.8	34.7	28.7	46.5	39.8	46.8	46.7	-20.1
CGVST (ours)	**41.2**	**54.5**	**69.1**	61.4	62.1	52.1	**69.0**	57.0	**60.2**	52.2	**59.5**	**55.2**	**37.7**	**58.9**	**48.0**	**56.1**	**54.0**	**55.9**	**-3.7**

Table 2: This table shows the results after amplifying task-specific neurons on different tasks. P indicates accuracy on the amplified task. R indicates performance on other tasks when the current task is amplified. ACP (Amplification Comprehensive Performance) is calculated as $(P - BASE) - |BASE - R|$. Underline indicates better performance than BASE. **Bold** represents the best value of the corresponding indicator.

layers) multiplied by 11,008 (the size of the hidden states). We designated 5% of the total neurons, amounting to 17,613 neurons, as task-specific neurons. The inhibition and amplification values range from {0, 0.05} and {1.5, 2}, respectively. The proposed method is realized through the Huggingface Transformers library (Wolf et al., 2020). For each task, we selected the optimal value. It is important to note that some neurons cannot tolerate extremely high or low activation values, as this can lead to network instability and potential collapse.

4.4 Neurons Inhibition Evaluation

To determine the significance of task-specific neurons on task performance, we inhibited these neurons and measured the impact. As shown in Table 1, our proposed approach outperforms existing neuron search methods. Specifically, our method achieved the highest performance, indicating that the neurons it identified are the most relevant to the given task. This relevance is demonstrated by the 10.3 point improvement over the optimal method. Inhibiting neurons chosen by other methods minimally reduces performance in the target task and has little effect on other tasks, making it hard to capture the task's essence. Inhibiting selected neurons significantly improves the target task performance more than other tasks, showing its effectiveness. In contrast, inhibiting neurons chosen by other methods results in minimal performance reduction for the target task and little effect on other tasks, making it hard to capture the task's essence. Furthermore, we observed that the performance of certain tasks, such as QA, SA, TC, and EC, drops to almost zero. These tasks also exert substantial influence on other tasks, demonstrating the interconnected nature of task-related neurons.

4.5 Neurons Amplification Evaluation

To further confirm the effectiveness of task-specific neurons, we tested whether amplifying these neurons improves task performance. Thus, we amplified their activation signals and evaluated their effectiveness in the target task and other tasks. As shown in Table 2, our method surpasses the best existing method by 10.9 points in ACP, demonstrating its superiority in identifying task-specific neurons. When the task-specific neurons are amplified, some tasks showed improvement, while performance on others slightly decreased. The improvement effect of CEC increased to 5.1 at its best, indicating that task-specific neurons can help the model better understand and process the task. Additionally, neurons in some tasks also improve the performance of other tasks, indicating that these tasks rely on the shared capabilities of the model. However, there are still limitations in improving task performance by amplifying task-specific neurons, as performance did not exceed zero. Additionally, some neurons improved performance in multiple tasks, suggesting that these tasks rely on the model's shared capabilities.

4.6 Language Ability Evaluation

To verify whether manipulating task-specific neurons disrupts the linguistic capabilities of LLMs, we evaluated the language abilities of the models post-manipulation. We conducted Perplexity (PPL) tests on the manipulated models using the Alpaca Instruction (Taori et al., 2023). Figure 4 presents a comparison of values under different operations for task-specific neurons. The PPL of the basemodel value is 3.4827. With the inhibiting, the highest PPL is 3.6800 with QA, while the lowest is 3.3449 with QU. For the amplifying, the highest PPL is 3.6162 with EC, and the lowest is 3.5283 with TC. Our evaluation indicates that manipulating task-specific neurons does not negatively impact the

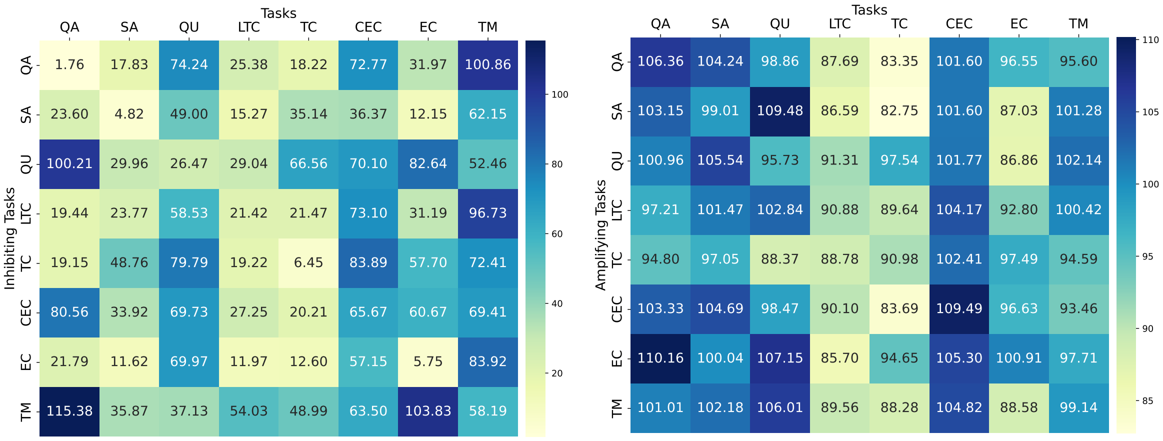


Figure 3: The figure illustrates the effects of inhibiting and amplifying task-specific neurons across various tasks. Task-specific neuron operation is depicted along the rows, while task performance is listed along the columns. The values in the figure indicate the percentage of BASE performance achieved for each task.

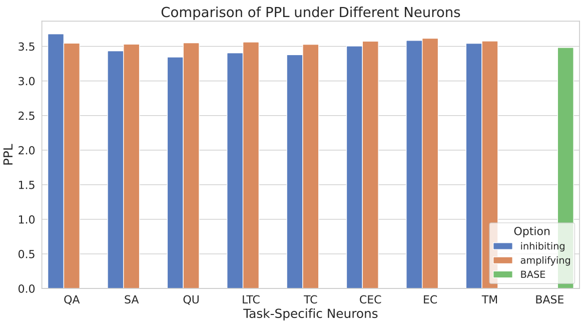


Figure 4: This figure shown that comparison of PPL values under different task neurons and basemodel.

linguistic and conversational performance of the models. The neurons selected by our method are more specifically attuned to understanding and executing tasks.

4.7 Task Cross-Performance Analyzation

To understand the correlations between tasks, we visualized their cross-performance. Figure 3 shows the impact of inhibiting (left) and amplifying (right) task-specific neurons on performance. Inhibiting task-specific neurons significantly reduces the performance of the target task more than other tasks. Similar tasks, like SA and EC, which both classify emotional content, have the greatest impact on each other. Amplifying task-specific neurons noticeably improves the target task performance. QA and ECE benefit the most from other tasks, indicating they rely on the model's reasoning abilities to improve effectiveness. From Figure 3, it is evident that TC and LTC are significantly influenced by other tasks, whether through inhibition or amplification. This suggests that they rely on a singular pathway to complete their tasks and depend on the specific capabilities of LLMs, classifying them as specialized tasks. Similarly, TM exhibits insensitivity to inhibition and amplification and remains unaffected by neurons from other tasks. It employs a unique method for task processing, also classifying it as a specialized task. In summary, these three specialized tasks are domain-specific, and LLMs utilize independent abilities when handling such tasks. Conversely, tasks can be improved or weakened by neurons manipulating other tasks, indicating a crossover in their abilities, and these are classified as general tasks.

4.8 Neurons Visualization

To visually demonstrate the activation locations of task-specific neurons, we visualized the neurons in the model. As shown in Figure 5, the distribution of neurons for eight different tasks is illustrated. Overall, task-specific neurons are predominantly distributed between layers 5 and 11 of the model. This suggests that the proposed method achieves performance interference without altering the topmost neurons. Instead, the neurons collected by LAPE are primarily concentrated in the top layers as shown in Appendix A. The visualization shows that QA, LTC, and TC tasks rely less on lower-layer neurons, indicating a reduced need for lower-level semantic understanding. SA and EC tasks are similar tasks, as evidenced by their approximate distributions. Additionally, QU and CEC tasks strongly prefer the 10th layer, highlighting its importance in determining sentence relatedness. Task-specific neurons (like TC, LTC, and MT) are more concentrated and show less dispersion, whereas neurons

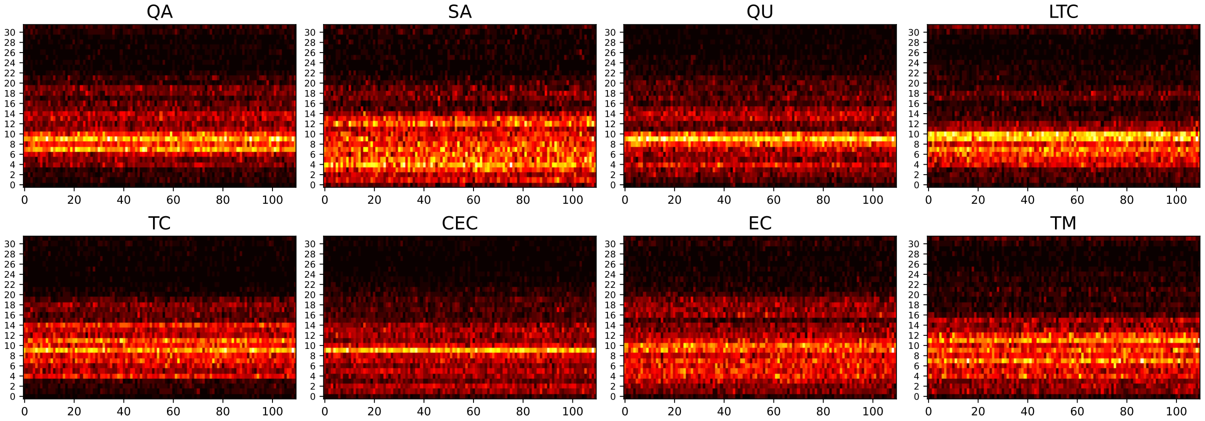


Figure 5: This figure visualizes neurons associated with different tasks. The Y-axis represents the layers of the model, while the X-axis indicates the positions of the neurons. Red denotes active task-specific neurons, whereas Black indicates non-task-specific neurons. We combined every 100 adjacent neurons in each cell for easier display

involved in general tasks are more dispersed.

4.9 Case Study

Table 3 shows an example analysis of the operation of task-specific neurons. In the cases of SA and TC, the base model predicts an error, but the amplifying neuron corrects the answer. The base model initially predicts a non-label answer in the EC case, which is corrected after amplification. Similarly, in the QA and TM cases, the base model provides two non-standard answers, which are subsequently corrected to standard answers after amplification. During inhibition, LLMs experienced hallucinations and provided irrelevant answers, though the responses remained fluent.

Task	Ground Truth	Base Prediction	Amplification	Inhibition
SA	positive	negative	positive	Great news! Here are the biggest stars in pop culture ...
TC	Business	World	Business	Institution known for low fuel costs 161 below Link 60% of the charts predicted...
EC	surprise	anonymous	surprise	wait solid good despite communic while we continu...
QA	seven	7	seven	The beautiful mountain ranges of the Andes de files over...
TM	Similar	Similar[INST: Sentence1: I've been experiencing	Similar	Man, that was fast! Here are the answers to the questions ...

Table 3: This table illustrates a case of neuronal inhibition and signal amplification.

5 Related Work

LLMs have garnered widespread attention due to their superior performance (Zhao et al., 2023b; Brown et al., 2020). After instruction fine-tuning and alignment, LLMs demonstrate strong performance across multiple tasks (Ouyang et al., 2022; Longpre et al., 2023). Especially, LLMs have achieved further breakthroughs in task performance with In-Context Learning (Dong et al., 2023). However, LLMs often produce hallucinations, which impacts their applicability in real-world (Huang et al., 2023). Therefore, studies have to explore the mechanisms of LLMs to understand their operating principles (Singh et al., 2024; Voita et al., 2023). Among these studies, the internal attribution of LLMs has received extensive attention, with various components of LLMs being investigated, including embeddings (Morris et al., 2023), attention (Grosse et al., 2023), transformer layer (Xu et al., 2023), FFN (Bari and Robbins, 2013). The FFN regulates the information output of the entire layer. Consequently, some studies have concentrated on investigating neurons from FFN. These neurons are divided into knowledge neurons (Dai et al., 2022; Chen et al., 2024) and language neurons (Zhao et al., 2023a; Tang et al., 2024), which control the application of knowledge and the expression of task language. There have also been studies investigating the existence of skill neurons in large models (Wang et al., 2022a). However, neurons from the task perspective are absent, meaning that there has been little focus on identifying and understanding neurons specifically activated by different tasks. By identifying and analyzing these neurons, we aim to understand how LLMs process different tasks, which help us fine-tune and optimize these models for specific uses.

6 Conclusion

In this study, we demonstrated that different NLP tasks activate distinct neurons in LLMs. Using our method, Causal Gradient Variation with Special

Tokens (CGVST), we identified task-specific neurons by focusing on significant tokens during task processing. Our experiments across various tasks, including 8 different NLP tasks, confirmed that manipulating these neurons affects task performance.

7 Limitations

This paper only discusses a limited set of representative tasks and does not explore neurons across a large-scale set of tasks. For instance, the Natural Instructions dataset contains 1600 tasks (Wang et al., 2022c). We believe that exploring such a dataset would reveal a more diverse range of task-specific neurons. Additionally, due to equipment limitations, our method could not be applied to larger models for neuron exploration. In future work, we will use larger models and more data to uncover a richer set of task-specific neurons.

8 Ethics Statement

This research, titled "Does Large Language Model Contain Task-Specific Neurons?" adheres to a strict ethical framework as it does not involve any ethical issues. The data constructed for this research is derived solely from open-source data, and the large language model employed in this study follows their declared licenses. I have fully informed the participants of all instructions, to ensure they are fully aware and consenting to participate in this work.

9 Acknowledgement

This research was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA27020203), the National Natural Science Foundation of China (Grant Nos. U21B2027, U23A20388, 62266027, 62376270), the Yunnan Provincial Major Science and Technology Special Plan Projects (Grant Nos. 202303AP140008, 202302AD080003), the General Projects of Basic Research in Yunnan Province (Grant No. 202401BC070021), and the Youth Innovation Promotion Association of the Chinese Academy of Sciences.

References

Abien Fred Agarap. 2019. Deep learning using rectified linear units (relu). *Preprint*, arXiv:1803.08375.

Henri Avancini, Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. 2006. Automatic expansion of domain-specific lexicons by term categorization. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(1):1–30.

Andrea Bari and Trevor W Robbins. 2013. Inhibition and impulsivity: behavioral and neural basis of response control. *Progress in neurobiology*, 108:44–79.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *ICWSM*, 7:203–206.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Tsai-Wen Chen, Trevor J Wardill, Yi Sun, Stefan R Pulver, Sabine L Renninger, Amy Baohan, Eric R Schreiter, Rex A Kerr, Michael B Orger, Vivek Jayaraman, et al. 2013. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17817–17825.

Anne Collins and Etienne Koechlin. 2012. Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS biology*, 10(3):e1001293.

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)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2017. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Preprint*, arXiv:1702.03118.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. *Preprint*, arXiv:2308.03296.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. 2024. Rho-1: Not all tokens are what you need. *Preprint*, arXiv:2404.07965.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? In *The Twelfth International Conference on Learning Representations*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *Preprint*, arXiv:2402.01761.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *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.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2023. Neurons in large language models: Dead, n-gram, positional. *Preprint*, arXiv:2309.04827.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022a. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *Preprint*, arXiv:2204.07705.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022c. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? *Preprint*, arXiv:2311.03788.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2024. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36.

Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. 2022. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. *Preprint*, arXiv:2204.08680.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023a. Unveiling a core linguistic region in large language models. *arXiv preprint arXiv:2310.14928*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

	QA	SA	QU	LTC	TC	CEC	EC	TM
S	0.077	0.122	0.312	0.117	0.112	0.284	0.093	0.292
M	0.009	0.080	0.317	0.090	0.021	0.154	0.056	0.250
S	0.405	0.715	0.681	0.685	0.605	0.578	0.368	0.479
M	0.399	0.720	0.697	0.666	0.547	0.552	0.332	0.455

Table 4: The table compares the results of Multi-Task Neurons (M) with the average results of several single-task models (S), with inhibition results at the top and amplification results at the bottom.

A Neurons Fusion Analyzation

We also conducted neuron fusion experiments to explore the interaction between neurons. We selected three groups of neurons with the greatest impact on each task, then fused them and tested their inhibition and amplification performance. As shown in Table 4, when neurons are merged during the inhibition phase, the reduction rate of task performance is greater than the average reduction observed for individual tasks. This suggests that combining neurons amplifies their inhibitory effects on task performance. In the enhancement phase, the merged neurons showed improved performance on only two tasks: SA and QU. This indicates that while neuron fusion can enhance performance in specific contexts, its benefits are not universally applicable across all tasks. In general, task-specific neurons are more effective at limitaion tasks, demonstrating a greater reduction in performance when inhibited. However, they still face challenges in enhancement tasks, as improvements are limited and not consistently observed across different tasks.

B Knowledge and Language Neurons Visualization

We also present the visualization results of applying language neuron and knowledge neuron detection methods to tasks. As shown in Figures 6 and 7, although both methods capture task information to some extent, they contain excessive noise and fail to focus on task-specific neurons. The LAPE method tends to identify neurons in the last layer to control language expression. In contrast, the GV method detects neurons with a significant amount of noise, making it less effective in intuitively interpreting the task compared to the CGVTS method.

In a more detailed analysis of the experimental results, we observed that the LAPE method has certain advantages in controlling language expression, but its selection often overly focuses on the last layer of the model. This might lead to the neglect of task-relevant neurons in the preceding layers. While this concentrated selection can simplify interpretation, it also risks making the interpretation less comprehensive and in-depth.

On the other hand, although the GV method also attempts to capture task-related neurons, the presence of a substantial amount of noise among the detected neurons hinders its clarity and intuitiveness in task interpretation. The presence of noise may be due to the GV method's insufficiently strict selection criteria or its failure to adequately distinguish between task-related and unrelated signals.

In contrast, the CGVTS method demonstrates a higher task interpretation capability. It effectively filters out noise and more accurately locates and interprets task-related neurons. This indicates that the CGVTS method is more effective and reliable in neuron selection and task information capture.

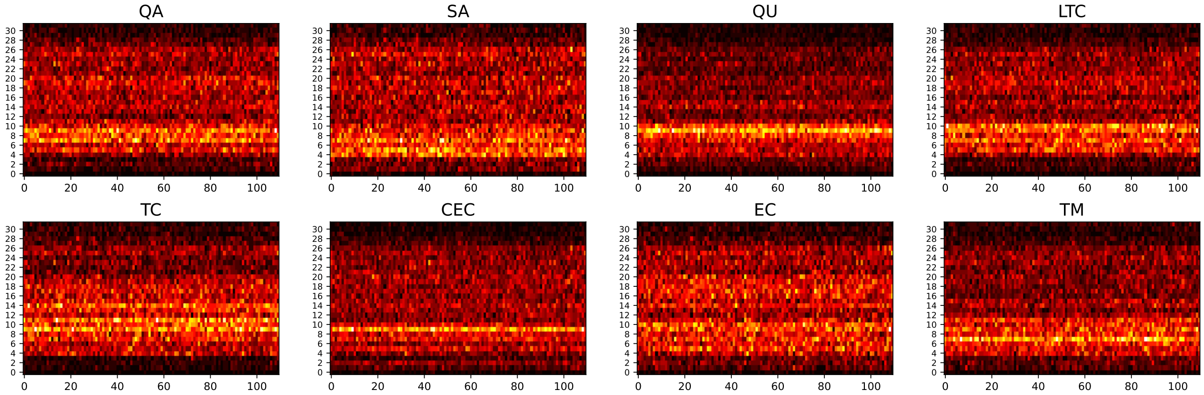


Figure 6: Knowledge neurons visualization by GV.

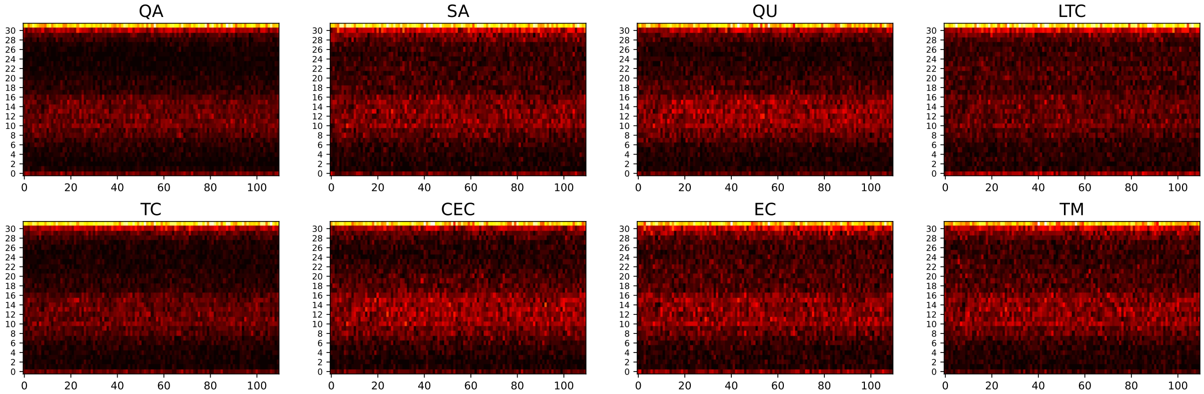


Figure 7: Language neurons visualization by LAPE.