Are Emotion and Rhetoric Neurons in LLM? Neuron Recognition and Adaptive Masking for Emotion-Rhetoric Prediction Steering

Li Zheng[1], **Xin Zhang**[2], **Shuyi He**[1], **Fei Li**[1]*, **Chong Teng**[1], **Jiangming Yang**[2], **Donghong Ji**[1]*, **Zhuang Li** [3]

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China

[2] Ant International [3]School of Computing Technologies, RMIT University, Australia

{zhengli,heshuyi354,lifei_csnlp,tengchong,dhji}@whu.edu.cn

{evan.zx,jmyang}@ant-intl.com, zhuang.li@rmit.edu.au

Abstract

Accurate comprehension and controllable generation of emotion and rhetoric are pivotal for enhancing the reasoning capabilities of large language models (LLMs). Existing studies mostly rely on external optimizations, lacking in-depth exploration of internal representation mechanisms, thus failing to achieve fine-grained steering at the neuron level. A handful of works on neurons are confined to emotions, neglecting rhetoric neurons and their intrinsic connections. Traditional neuron masking also exhibits counterintuitive phenomena, making reliable verification of neuron functionality infeasible. To address these issues, we systematically investigate the neurons representation mechanisms and inherent associations of 6 emotion categories and 4 core rhetorical devices. We propose a neuron identification framework that integrates multi-dimensional screening, and design an adaptive masking method incorporating dynamic filtering, attenuation masking, and feedback optimization, enabling reliable causal validation of neuron functionality. Through neuron regulation, we achieve directed induction of non-target sentences and enhancement of emotion tasks via rhetoric neurons. Experiments on 5 commonly used datasets validate the effectiveness of our method, providing a novel paradigm for the fine-grained steering of emotion and rhetoric expressions in LLMs.

1 Introduction

Emotion and rhetoric are fundamental components of human communication (Konstan, 2007; Fussell and Moss, 2014; Zheng et al., 2025c,d). Emotions convey speakers' subjective attitudes and affective states, while rhetorical devices shape how meanings are intensified, softened, or implied. As large language models (LLMs) are increasingly deployed in intelligent dialogue (Ou et al., 2024; Xu et al., 2024; Zheng et al., 2025b), creative writing (Wu et al., 2025; Zhong et al., 2024), and customer service (Su et al., 2025; Li et al., 2025), robust understanding and controllable expression of emotion and rhetoric become essential for improving user experience, safety, and reliability in real-world interactions (Hu et al., 2017).

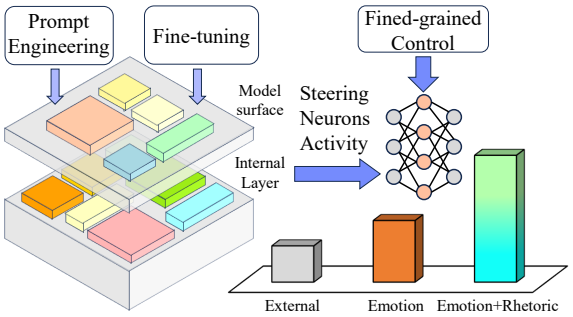

Figure 1: Emotion recognition accuracy comparison on DailyDialogue (Li et al., 2017): external optimization vs. emotion and rhetoric neurons steering. "External" denotes prompt engineering or fine-tuning. "Emotion" indicates the injection of emotion neurons. "Emotion+Rhetoric" denotes the injection of emotion neurons augmented with rhetoric neurons.

As illustrated in Figure 1, most existing approaches improve emotion- and rhetoric-related performance primarily through *external* optimization, such as prompt engineering (Brown et al., 2020; Zheng et al., 2024) or fine-tuning (Liu et al., 2024). While effective at the output level, these methods provide limited causal insight into the internal representations that support emotion and rhetoric, and they often offer weak fine-grained controllability over specific affective or rhetorical attributes. Meanwhile, recent interpretability studies have started to investigate *emotion-selective neurons* (Lee et al., 2025; Di Palma et al., 2025); however, two key gaps remain. On the one hand, the internal mechanisms of *rhetorical* representations, and more importantly, the relationship between emotion and rhetoric inside LLMs, are still underexplored, leaving unclear whether and how rhetorical signals modulate emotional representa-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4206–4216

tions. On the other hand, widely used neuron-level *function verification* practices (e.g., forced-zero ablation or mean substitution) can behave unreliably on these tasks: masking neurons that appear highly related to a target label does not necessarily lead to the expected performance degradation, and can even yield counterintuitive changes. This raises a practical challenge: without a reliable intervention scheme, it is difficult to make trustworthy causal claims about neuron functionality, let alone use such neurons for controllable manipulation.

Motivated by these gaps, we systematically study neuron-level representations of emotion and rhetoric, as well as their interactions, covering 6 emotion categories (happiness (hap), sadness (sad), anger (ang), fear (fea), surprise (sup), and disgust (dig)) and 4 rhetorical devices (metaphor (met), hyperbole (hpy), humor (hum), and sarcasm (sar)). Specifically, we ask:

> *Can we reliably localize neurons that are selective to specific emotion and rhetoric labels, verify their causal roles through stable masking interventions, and leverage them for controllable steering, including probing potential emotion–rhetoric interactions?*

To answer this question, we develop a neuron-level analysis and intervention framework. **First**, we identify candidate emotion and rhetoric neurons by combining activation frequency statistics, probability normalization, and entropy-based selectivity filtering, and we characterize their layer-wise distributions. **Second**, to address the instability and counterintuitive behaviors of traditional masking, we propose an *adaptive masking* strategy that integrates dynamic neuron selection, attention-based masking, and feedback-driven adjustment. This design aims to produce a consistent decline in task performance when masking truly relevant neurons, improving the reliability of causal verification. **Third**, we demonstrate controllable manipulation by intervening on the identified neuron sets to steer predictions from non-target to target emotion/rhetoric categories. Furthermore, we study cross-signal interactions by injecting rhetoric-neuron signals into emotion recognition, testing whether rhetorical representations can assist emotional discrimination.

We conduct extensive experiments on five widely used emotion and rhetoric datasets, namely DailyDialogue (Li et al., 2017), HYPO (Troiano et al., 2018), TroFi (Birke and Sarkar, 2006), IAC-v2 (Oraby et al., 2016), and ColBERT (Annamoradnejad and Zoghi, 2024), from which we derive the following insights. ❶ Emotion and rhetoric neurons tend to exhibit stronger activation in upper layers. ❷ Rhetoric-neuron activations can improve the separability of emotional features, benefiting emotion recognition. ❸ The proposed adaptive masking yields a stable performance decline after masking, enabling more reliable function verification than standard ablation/substitution baselines. ❹ By manipulating the identified neurons, predictions for non-target emotion/rhetoric inputs can be directionally induced toward target categories.

Our main contributions are as follows:

- We conduct the first systematic investigation into the foundations of rhetoric neurons and confirm the auxiliary role of rhetoric in emotion recognition.
- We propose an adaptive neuron masking method that enables reliable causal verification of neuron functions.
- We achieve controllable manipulation of emotion and rhetoric neurons, offering a novel pathway for the fine-grained steering of emotional and rhetorical expressions in LLMs.

2 Observation and Key Intuition

Most existing approaches improve emotion- and rhetoric-related performance primarily via output-level optimization (e.g., prompting or fine-tuning), which provides limited neuron-level causal evidence and offers weak fine-grained controllability. Moreover, when we attempt to verify neuron functionality using standard masking interventions (e.g., forced-zero or mean substitution), we observe counterintuitive behaviors on emotion and rhetoric tasks (Figure 2), making causal verification unreliable in this setting. Motivated by these observations, we develop an intervention-centric framework that localizes label-selective neurons and employs more stable masking and activation-based interventions for verification and steering.

2.1 Observation on Existing Limitations

Limited neuron-level insight and steering. Output-level optimization methods (e.g., prompt engineering and fine-tuning) are effective for improving task performance, but they do not directly provide causal evidence about which internal representations are responsible for emotion and rhetoric

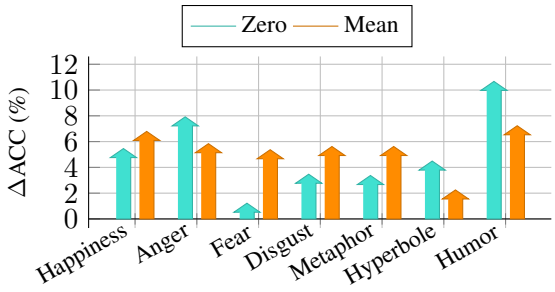

Figure 2: **Observation**. Performance changes of emotion and rhetoric tasks under traditional target neuron masking methods.

predictions, nor do they enable targeted neuron-level steering. In addition, the potential interaction between emotion and rhetoric representations inside LLMs is underexplored, limiting opportunities to leverage rhetorical cues to assist emotion recognition and to build more controllable systems. Existing methods (Kim et al., 2025; Yang et al., 2025b) fail to recognize this core mechanism, and only optimize the model output through surface-level text features, resulting in two major problems. One is the lack of controllability and synergism. Precise modulation of emotional or rhetorical expressions cannot be achieved, and the failure to explore the correlation between emotion and rhetoric results in the absence of rhetorical cues for emotional recognition tasks. The other is insufficient interpretability, such that performance improvements cannot be traced back to the causal mechanism of neuronal functioning.

Unreliable verification under standard masking. Masking-based methods aim to identify task-relevant hidden features by intervening on internal representations (Feng et al., 2024), yet standard interventions can behave counterintuitively on emotion and rhetoric tasks. As shown in Figure 2, forcing selected neurons to zero or substituting them with a mean value does not consistently reduce accuracy and can even increase it. We hypothesize that such behaviors are consistent with redundancy or redistribution effects (e.g., alternative pathways becoming more dominant under hard ablation), and that mean substitution may leave residual information that still supports inference. These observations motivate a more stable intervention scheme for function verification.

2.2 Key Intuition of Our Method

Manipulation of internal neurons. The emotion and rhetoric neurons exhibit functional specificity and activation stability. Specific neurons only respond strongly to a certain type of emotion or rhetoric, and this response pattern remains consistent across different scenarios. More importantly, a synergistic activation effect exists between the two neurons. The activation of rhetoric neurons will significantly enhance the response intensity of the corresponding emotion neurons, providing a basis for the associated regulation of emotion and rhetoric. As illustrated in Figure 1, the performance of relying solely on external optimization is significantly weaker than that of directly regulating neurons. Notably, co-regulating emotion and rhetoric neurons yields superior task performance.

Adaptive masking methods. The counterintuitive phenomenon observed in traditional masking methods stems essentially from a mismatch between the masking strategies and the functional characteristics of neurons. Thus, the key to successful masking lies in identifying genuinely core neurons via activation discrepancy screening, followed by intervening in their functions through attenuation masking instead of complete ablation. Attenuation masking not only avoids triggering the model's functional compensation mechanism but also precisely weakens the representational capacity of target neurons, leading to a steady decline in task accuracy and enabling reliable validation of their functions.

Takeaway. *We provide a systematic study of rhetoric-selective neurons and their interaction with emotion-selective neurons. We further propose an adaptive, attenuation-based masking scheme that yields more stable verification behavior than standard masking baselines, and demonstrate neuron-level steering of emotion and rhetoric predictions via activation interventions.*

3 Methodology

As illustrated in Figure 3, our framework adopts the Llama-3.1-8B-Instruct as the research vehicle. This model is built on a Transformer decoder architecture, comprising 32 Transformer blocks where each block consists of a multi-head self-attention (MHA) module and a feed-forward network (FFN) module. Existing research (Lee et al., 2025) has confirmed that FFN layers serve as the core representation region for semantic features, thus we focus on the neurons within FFN layers.

3.1 Neuron Recognition

Neuron activation mechanism. After the MHA module, the FFN layer performs additional non-

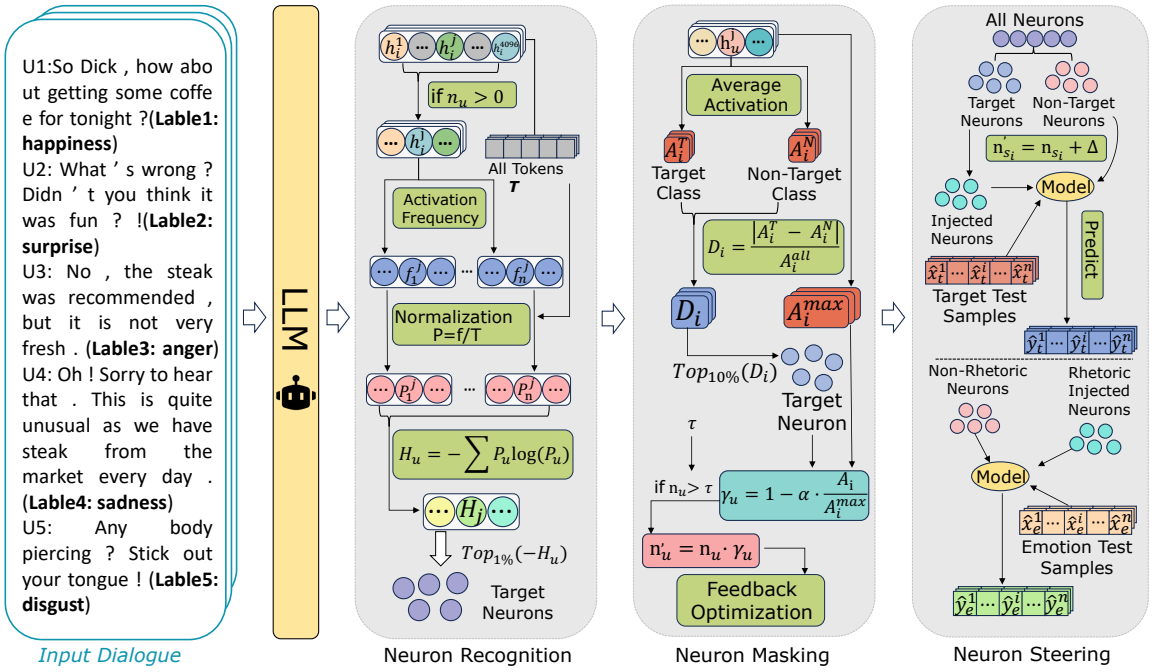

Figure 3: The overall architecture of our framework. h_i^j denotes the representation of the j-th neuron in the i-th layer. The target class refers to a specific emotion or rhetoric category subject to manipulation.

linear transformations on token features to further distill semantic information.

$$\text{FFN}'(X_{\text{mid}}) = g(X_{\text{mid}} W_{\text{in}}) W_{\text{out}} \tag{1}$$

where X_{mid} denotes the output features of the MHA module, W_{in} and W_{out} represent the input and output weight matrices of the FFN layer, and $g(\cdot)$ is the activation function. For a neuron u within the FFN layer, its activation condition is defined as:

$$n_u = \max(0, h_u) \tag{2}$$

where h_u is the output of neuron u after linear transformation. Neuron u is determined to be activated when $n_u > 0$.

Recognizing neurons. We first calculate the activation frequency. For each neuron n_u in the FFN layer of each model layer, we count its activation times when inputting sentences corresponding to a emotion label e or rhetorical label r, denoted as $f_{u,e}$ or $f_{u,r}$. Then, we perform activation probability normalization. Let T_e (or T_r) be the total number of tokens corresponding to emotion e or rhetoric r. The activation probability $P_{u,e}$ or $P_{u,r}$ is given by:

$$P_{u,e} = \frac{f_{u,e}}{T_e}, P_{u,r} = \frac{f_{u,r}}{T_r} \tag{3}$$

Finally, we use entropy to measure the concentration of this probability distribution. A lower entropy indicates that the neuron's responses are more concentrated on specific emotions or rhetorics. The entropy is calculated as follows:

$$H_{e/r} = -\sum_{e/r \in E/R} P_{u,e/r} \log(P_{u,e/r}) \tag{4}$$

where E denotes the set of 6 emotion categories, and R denotes the set of 4 rhetorical categories. We select the top 1% neurons with the lowest entropy as emotion or rhetoric neurons.

3.2 Neuron Masking

Traditional masking methods. Traditional masking methods verify functionality by directly intervening in neuron activation values, encompassing two classical implementation paradigms. One is the forced zero method, which directly resets the activation value of the target neuron to 0, with the formula given as follows:

$$n'_u = 0 \tag{5}$$

where n'_u denotes the activation value of the neuron after masking.

The other is the mean substitution method, which replaces the original activation value of the target neuron with the global mean of the activation values of all non-target neurons in the FFN layer where the target neuron resides.

$$n'_u = \frac{1}{N - M} \sum_{k \in \Omega} n_k \tag{6}$$

where N is the total number of neurons in the FFN layer, M is the number of target neurons, Ω represents the set of all non-target neurons in the layer, and n_k is the original activation value of the k-th non-target neuron in Ω.

Adaptive masking methods. To address the counterintuitive phenomenon of traditional masking methods in emotion and rhetorical tasks, we propose an adaptive masking method integrating dynamic selection and feedback optimization. Throughout our experiments, the LLM parameters are kept fixed; we intervene only on FFN activations. The feedback optimization (adjusting the selection criterion and α) is performed on a held-out development split only; test sets are used once for final reporting without further updates.

First, we collect neuron activation patterns. We feed the data into the model, and record the average activation value of each neuron i in each FFN layer for the two types of sentences, denoted as $A_{i,\text{target}}$ (average activation for target sentences) and $A_{i,\text{non-target}}$ (average activation for non-target sentences), respectively. On this basis, we define the activation difference D_i to quantify the response specificity of neuron i:

$$D_i = \frac{|A_{i,target} - A_{i,non-target}|}{A_{all}} \tag{7}$$

where A_{all} denotes the average activation for all sentences. A larger D_i indicates that neuron i possesses a stronger ability to discriminate between target and non-target sentences.

Subsequently, we perform core neuron selection. By setting an activation threshold τ , we select the top 10% of neurons ranked by D_i to construct a set S of target neurons that are significantly activated only in target sentences, ensuring that subsequent masking operations focus on critical neurons. Then we perform dynamic masking by applying attenuated masking to neurons in set S. The activation value of the neuron after masking is as follows:

$$n'_u = n_u \times (1 - \alpha \times \frac{A_i}{A_i^{max}}) \tag{8}$$

where α denotes the attenuation coefficient, which is used to flexibly adjust the masking intensity.

To further ensure the reliability of the masking effect, we introduce a feedback optimization mechanism for iterative adjustment. The first step is accuracy monitoring. We calculate the task accuracy $\text{Acc}_{\text{adapt}}$ after adaptive masking and conduct a quantitative comparison with the original accuracy $\text{Acc}_{\text{origin}}$ before masking. The second step is dynamic parameter adjustment. If $\text{Acc}_{\text{adapt}} \geq \text{Acc}_{\text{origin}}$, we increase the core neuron selection threshold τ and the attenuation coefficient α to enhance the intensity of the masking operation.

3.3 Neuron Steering

The core objective of controllable neuronal manipulation is to achieve the controllable output of emotions and rhetorical devices by precisely regulating the activation states of emotion and rhetoric neurons. Concurrently, we investigate the intrinsic correlation between emotion and rhetoric, and validate the enhancing effect of rhetoric neurons on emotion recognition tasks.

Rhetoric and emotion neuron steering. Based on the identified set of core target neurons, we extract the activation values of the neurons within this set to construct a functional vector $\mathbf{V} = [n_{s_1}, n_{s_2}, ..., n_{s_k}]$. k denotes the number of neurons in the core neuron set, and n_{s_i} represents the original activation value of the i-th neuron in the set. When a non-target-type sentence is input, we apply activation to all neurons in the core target neuron set within the FFN layer of the model to induce the model to generate a target-type output.

$$n'_{s_i} = n_{s_i} + \beta \times \overline{n}_{s_i} \tag{9}$$

where n'_{s_i} denotes the modulated activation value of the neuron s_i, n_{s_i} is the original activation value of this neuron when a non-target sentence is input, $\overline{n}_{si}$ represents the average activation value of the neuron s_i under the input scenario of target sentences, and β is the activation intensity coefficient.

Finally, we validate the manipulation effect by comparing the model's classification results of non-target sentences before and after manipulation. If the model classifies sentences originally predicted as the non-target category into the target category after manipulation, it demonstrates the effectiveness of the neuronal manipulation.

Rhetoric neurons assisted emotion recognition. To validate the auxiliary enhancement effect of rhetoric on emotion recognition, we take metaphor as an example for illustration. First, we extract metaphor-selective neurons in a given FFN layer (width d), represented as an index set $\mathcal{I}_{\text{meta}} \subseteq \{1, \ldots, d\}$. For each neuron index $i \in \mathcal{I}_{\text{meta}}$, we record its average activation under metaphorical

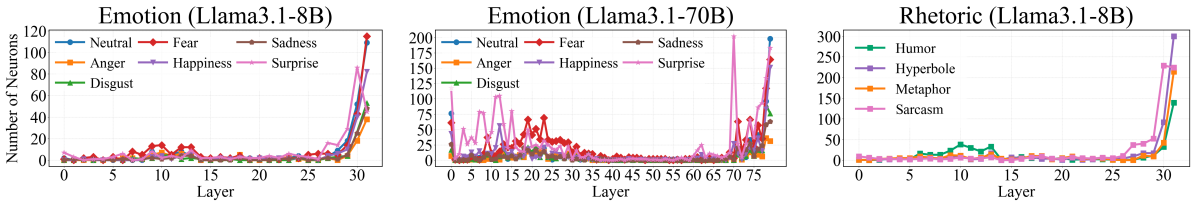

Figure 4: Distribution of emotion and rhetoric neurons.

Method	Emotion						Rhetoric			
	Happiness	Sadness	Anger	Fear	Surprise	Disgust	Metaphor	Hyperbole	Humor	Sarcasm
Zero	+5.46	+4.32	+7.92	+1.22	-2.85	+3.47	+3.37	+4.49	+10.67	-4.17
Mean	+6.79	-1.13	+5.84	+5.37	-3.59	+5.62	+5.62	+2.25	+7.23	-5.78
Adaptive	**-9.25**	**-8.63**	**-10.14**	**-7.85**	**-11.42**	**-14.26**	**-7.29**	**-13.48**	**-4.28**	**-10.61**
Cross-Dataset										
Zero	+3.24	+5.11	+4.39	+2.37	-3.72	+1.68	+5.64	+5.28	+9.31	-3.95
Mean	+4.61	-1.18	+4.79	+5.28	-4.11	+4.65	+6.14	+3.71	+7.58	-5.73
Adaptive	**-7.31**	**-6.83**	**-10.25**	**-6.37**	**-12.39**	**-11.65**	**-6.79**	**-14.27**	**-5.96**	**-11.63**

Table 1: Comparison results of different masking methods (accuracy change ΔACC (%)).

inputs, calculated as follows:

$$\overline{a}_{i,\text{meta}} = \frac{1}{T_{\text{meta}}} \sum_{t=1}^{T_{\text{meta}}} a_{i,t}, \tag{10}$$

where $a_{i,t}$ denotes the activation of the i-th FFN neuron in the t-th metaphor sample, and T_{meta} is the number of metaphor samples. We store these values as a metaphor feature library $\mathcal{F}_{\text{meta}} = \{\overline{a}_{i,\text{meta}}\}_{i \in \mathcal{I}_{\text{meta}}}$.

Then, when performing the emotion recognition task, we extract emotion-selective neurons as another index set $\mathcal{I}_{\text{emo}} \subseteq \{1, \ldots, d\}$ in the same FFN layer, and denote the activation of neuron i as a_i. We fuse neuron activations by element-wise injection. The final fused activation value of the emotion neuron is calculated as follows:

$$a_{i,\text{joint}} = a_i + \omega \cdot \overline{a}_{i,\text{meta}}, \quad \forall i \in \mathcal{I}_{\text{emo}}, \tag{11}$$

where $\omega \in [0, 1]$ controls the injection strength. Note that the index i refers to the same hidden unit in the same FFN layer for both a_i and $\overline{a}_{i,\text{meta}}$.

4 Experiments

4.1 Experimental Setting

We conduct evaluations on five widely used emotion and rhetoric datasets, namely DailyDialogue (emotion) (Li et al., 2017), TroFi (metaphor) (Birke and Sarkar, 2006), HYPO (hyperbole) (Troiano et al., 2018), ColBERT (humor) (Annamoradnejad and Zoghi, 2024), and IAC-v2 (sarcasm) (Oraby et al., 2016). In terms of evaluation metrics, we adopt Acc to assess the model's performance.

4.2 Localization and Distribution of Emotion and Rhetoric Neurons

In Figure 4, we visualize the layer-wise distribution of the *localized* emotion- and rhetoric-selective neurons and analyze their activation patterns across model layers. For Llama-3.1-8B, both emotion and rhetoric neurons exhibit distinct top-layer aggregation: activation remains consistently low across the first 25 layers, surges sharply, and peaks at the top layers. This distribution suggests the model relies more on top-layer semantic integration to represent and process emotions and rhetoric. In contrast, emotion neurons in Llama-3.1-70B show dual-end concentration, with high activation in both top and bottom layers. This indicates that as model scale increases, emotional representation is no longer confined to upper-layer semantic integration, bottom layers also participate in early capture and initial processing of emotion information.

4.3 Impact of Different Masking Methods

Setting. To validate the functional relevance of the localized target neurons, we keep the LLM parameters frozen and evaluate task accuracy under activation-level masking interventions. We optimize the adaptive masking policy (including the core-neuron selection threshold and attenuation co-

efficient α) on the development split of the original dataset, and report results on the original test set and five cross-datasets (IEMOCAP (Busso et al., 2008), LCC (Mohler et al., 2016), HYPO-L (Zhang and Wan, 2022), FunLines (Hossain et al., 2020), MUStARD (Castro et al., 2019)). We compare three masking methods: Zero masking (forced zeroing), Mean masking (mean substitution), and our adaptive masking.

Analysis. Experimental results in Table 1 show that adaptive masking consistently reduces accuracy on both the original test set and cross-datasets, providing evidence that the localized neuron sets are causally relevant to model decisions under our intervention scheme. In contrast, Zero masking exhibits a counterintuitive trend where accuracy increases post-masking, which we conjecture stems from functionally complementary neuron clusters within the model. Mean masking fails to induce notable accuracy drops in most tasks: mean substitution does not truly disrupt core neurons' specific functional representations, leaving residual semantic encoding intact and enabling the model to infer via residual features.

4.4 Rhetoric Neurons Assisted Emotion Recognition

We inject rhetoric neurons into emotion texts to investigate their auxiliary enhancement on emotion recognition, with results in Figure 5. Four typical rhetorical types exert positive effects on most emotion categories, verifying the potential of rhetoric neurons to boost emotion recognition performance. Metaphor neurons yield particularly prominent improvements for fear, as metaphors strengthen perception of emotions via concrete expressions. Hyperbole neurons exert positive effects across all emotions, aligning with their inherent trait of amplifying emotion intensity. Humor neurons negatively impact emotions like happiness and surprise, as humor's playful tone dilutes their original emotion concentration, but still benefit some emotions (e.g., anger, fear). Sarcasm neurons most significantly promote sadness, attributed to expressive compatibility between sarcasm's implicit criticism and sadness's restrained traits, enhancing the model's recognition of this emotion.

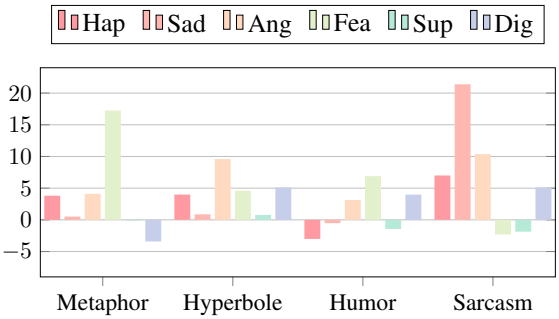

Figure 5: Experimental results of injecting rhetoric neurons into emotion recognition task.

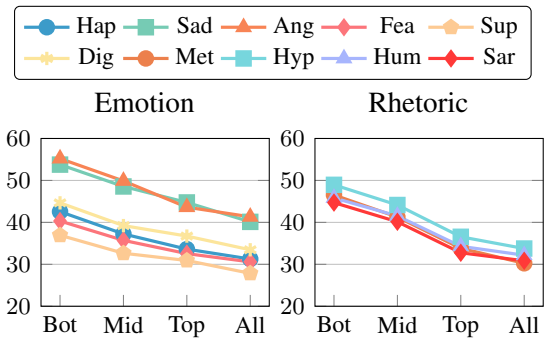

Figure 6: Comparison results of different layers of masking.

4.5 Impact of Masking Across Different Layers

In Figure 6, we explore the impact of neuron masking across different layers on task prediction outcomes. We find that all layer masking induces the most significant performance degradation across all emotion and rhetoric tasks. This indicates that the functional representations of emotion and rhetoric neurons are not confined to a single layer but instead rely on cross-layer synergistic interactions. Beyond all layer masking, top layer masking exerts the most pronounced performance impairment on emotion and rhetoric tasks, outperforming mid and bottom layer masking in terms of prediction accuracy reduction. This observation aligns with the distribution characteristic that emotion and rhetoric neurons exhibit higher activation intensity in the top layers. Since the top layers aggregate a larger number of emotion and rhetoric neurons, masking these layers deprives the model of effective semantic encoding support for emotion and rhetoric. This further validates the hierarchical distribution characteristics and functional relevance of emotion and rhetoric neurons in the model's architecture.

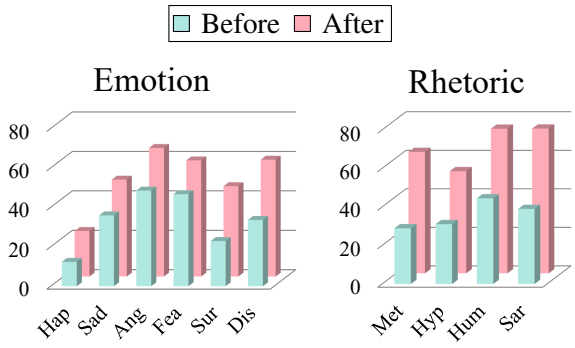

Figure 7: Comparison results before and after neurons manipulation.

4.6 Neuron-level Steering of Emotion and Rhetoric Predictions

We validate the controllable manipulation capability of emotion and rhetoric neurons in Figure 7. Specifically, we inject the functional features of target neurons into non-target sentences, thereby inducing the model to convert its predictions of non-target types into target types. The manipulation efficacy is quantified as the ratio of samples predicted as the target class to the total number of samples in the dataset. A higher ratio indicates that more non-target instances are induced to be classified as the target class, thus reflecting a stronger controllable manipulation capability. When manipulating emotion neurons, the coverage rate of all emotion categories is significantly improved after the injection of target neuron features. This indicates that modulation of the activation states of emotion neurons can effectively enhance the model's perception of target emotion features. Similarly, the controllability verification yields excellent results in the rhetoric neuron manipulation task. This demonstrates that the injection of functional features of rhetoric neurons can precisely guide the model to capture the semantic expression patterns of target rhetorical devices. It also successfully induces non-target rhetoric sentences to be classified as the target rhetoric category.

5 Related Work

5.1 Emotion Neurons

The exploration of emotion neurons aims to uncover the internal representation mechanisms of emotional information in LLMs, providing theoretical support for the fine-grained steering of emotion understanding and generation (Smith and Kornelsen, 2011; Jayasinghe et al., 2025; Wang et al., 2025; Zheng et al., 2025a). Existing studies (Liao et al., 2025; Huangfu et al., 2025) mostly focus on optimizing external emotion-related tasks. With the advancement of interpretability research, the exploration of internal mechanisms at the neuronal level has gradually emerged as a research hotspot (Zhang et al., 2025b; Tak et al., 2025). The concept of emotion neurons originates from Radford et al. (2018), referring to a set of neurons in the model that exhibit selective responses to specific emotions. Lee et al. (2025) validate the existence of emotion neurons in Llama-series models, finding that their distribution varies with model scale. Di Palma et al. (2025) demonstrate that linear classifiers can achieve high-accuracy emotion recognition via probing techniques.

5.2 Rhetorical Models

Research on rhetorical processing revolves around two core tasks: recognition (Yang et al., 2025a; Cocchieri et al., 2025; Zhang et al., 2025a) and generation (Stowe et al., 2021; Zhong et al., 2024; Goel et al., 2025). In rhetorical recognition tasks, feature engineering and fine-tuning of pre-trained models are the mainstream technical approaches. Saravia et al. (2018) achieve rhetoric type classification using contextual semantic features yet lack in-depth analysis of the underlying encoding mechanisms of rhetorical information in the model. Rajakumar and Boicu (2025) demonstrate the advantage of lightweight models in balancing efficiency and performance for practical deployment, but they do not provide an explanation for the internal operational principles of rhetorical representations. In rhetorical generation tasks, Deng et al. (2023) propose prompt reformulation to optimize the model's comprehension of ambiguous queries. Benara et al. (2024) introduce a question-answering embedding approach that enables modeling of semantic representations associated with rhetorical expressions.

6 Conclusion

In this paper, we systematically explore emotion and rhetoric neurons in LLMs, addressing key gaps: inadequate exploration of rhetorical neurons, ambiguous emotion-rhetoric associations, and unreliable causal validation with traditional masking methods. Via a synergistic framework integrating neuron recognition, adaptive causal validation masking, and controllable causal intervention, we uncover their distribution and functional mechanisms, establishing a reliable system for causal

validation and regulation. Experiments show our adaptive masking resolves counterintuitive drawbacks of traditional methods, yielding a robust tool for neuronal causal attribution. Neuro modulation enables directed induction of emotional and rhetorical outputs and emotion-task performance gains via rhetoric neurons, laying a theoretical and technical foundation for fine-grained LLM steering.

Limitations

Despite uncovering the internal neural mechanisms underlying emotion and rhetoric processing and advancing the interpretability of Large Language Models (LLMs), this study is not without limitations. First, the research scope is confined to 6 basic emotion categories and 4 core rhetorical devices; future work may extend to complex emotional states and diverse rhetorical forms. Second, neuronal manipulation in the current framework relies on static functional vectors, so context-aware dynamic adjustment strategies could further enhance manipulation precision.

Acknowledgments

This work was supported by Ant Group through the CCF-Ant Research Fund. Fei Li and Donghong Ji are co-corresponding authors.

References

Issa Annamoradnejad and Gohar Zoghi. 2024. Colbert: Using bert sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685.

Vinamra Benara, Chandan Singh, John X Morris, Richard J Antonello, Ion Stoica, Alexander G Huth, and Jianfeng Gao. 2024. Crafting interpretable embeddings for language neuroscience by asking llms questions. *Advances in neural information processing systems*, 37:124137.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European chapter of the association for computational linguistics*, pages 329–336.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629.

Alessio Cocchieri, Luca Ragazzi, Paolo Italiani, Giuseppe Tagliavini, and Gianluca Moro. 2025. "what do you call a dog that is incontrovertibly true? dogma": Testing llm generalization through humor. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22922–22937.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Llamas have feelings too: Unveiling sentiment and emotion representations in llama models through probing. *arXiv preprint arXiv:2505.16491*.

Tao Feng, Lizhen Qu, Zhuang Li, Haolan Zhan, Yuncheng Hua, and Reza Haf. 2024. IMO: Greedy layer-wise sparse representation learning for out-of-distribution text classification with pre-trained models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2625–2639, Bangkok, Thailand. Association for Computational Linguistics.

Susan R Fussell and Mallie M Moss. 2014. Figurative language in emotional communication. In *Social and cognitive approaches to interpersonal communication*, pages 113–141. Psychology Press.

Palaash Goel, Dushyant Singh Chauhan, and Md Shad Akhtar. 2025. Target-augmented shared fusion-based multimodal sarcasm explanation generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8480–8493.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020. Stimulating creativity with funlines: A case study of humor generation in headlines. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 256–262.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Yuanxiang Huangfu, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2025. Non-emotion-centric empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 989–999.

Hiruni Maleesa Jayasinghe, Kok Wai Wong, and Anupiya Nugaliyadde. 2025. A systematic review of interpretability and explainability for speech emotion features in automatic speech emotion recognition. *Pattern recognition*, page 112122.

Minseo Kim, Taemin Kim, Thu Hoang Anh Vo, Yugyeong Jung, and Uichin Lee. 2025. Exploring modular prompt design for emotion and mental health recognition. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

David Konstan. 2007. Rhetoric and emotion. *A companion to Greek rhetoric*, 411:25.

Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. 2025. Do large language models have "emotion neurons"? investigating the existence and role. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15617–15639.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zhipeng Li, Binglin Wu, Yingyi Zhang, Xianneng Li, Kai Li, and Weizhi Chen. 2025. Cusmer: Multimodal intent recognition in customer service via data augment and llm merge. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 3058–3062.

Jian Liao, Yu Feng, Yujin Zheng, Jun Zhao, Suge Wang, and Jianxing Zheng. 2025. My words imply your opinion: Reader agent-based propagation enhancement for personalized implicit emotion analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16156–16172.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the lcc metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 31–41.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2018. Learning to generate reviews and discovering sentiment.

Rohit Rajakumar and Mihai Boicu. 2025. Evaluating lightweight transformer models for rhetorical element classification in student essays. *Journal of Student-Scientists' Research*, 7.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697.

Stephen D Smith and Jennifer Kornelsen. 2011. Emotion-dependent responses in spinal cord neurons: a spinal fmri study. *NeuroImage*, 58(1):269–274.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736.

Hanchen Su, Wei Luo, Yashar Mehdad, Wei Han, Elaine Liu, Wayne Zhang, Mia Zhao, and Joy Zhang. 2025. Llm-friendly knowledge representation for customer support. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 496–504.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120. Association for Computational Linguistics.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304.

Yuchi Wang, Junliang Guo, Jianhong Bai, Runyi Yu, Tianyu He, Xu Tan, Xu Sun, and Jiang Bian. 2025. Instructavatar: Text-guided emotion and motion control for avatar generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8132–8140.

Zhikun Wu, Thomas Weber, and Florian Müller. 2025. One does not simply meme alone: Evaluating co-creativity between llms and humans in the generation of humor. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1082–1092.

Zhenyu Xu, Hailin Xu, Zhouyang Lu, Yingying Zhao, Rui Zhu, Yujiang Wang, Mingzhi Dong, Yuhu Chang, Qin Lv, Robert P Dick, et al. 2024. Can large language models be good companions? an llm-based eyewear system with conversational common ground. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–41.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Jenny Zhang, Yiliao Song, Hongfei Lin, and Feng Xia. 2025a. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26301–26317.

Yang Yang, Xunde Dong, and Yupeng Qiang. 2025b. Mse-adapter: A lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25642–25650.

Jiecheng Zhang, CL Philip Chen, Shuzhen Li, and Tong Zhang. 2025a. Incongruity-aware tension field network for multi-modal sarcasm detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14499–14508.

Xinjie Zhang, Tenggan Zhang, Lei Sun, Jinming Zhao, and Qin Jin. 2025b. Exploring interpretability in deep learning for affective computing: a comprehensive review. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Yunxiang Zhang and Xiaojun Wan. 2022. Mover: Mask, over-generate and rank for hyperbole generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6018–6030.

Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. 2025a. Multi-granular multimodal clue fusion for meme understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26057–26065.

Li Zheng, Hao Fei, Fei Li, Bobo Li, Lizi Liao, Donghong Ji, and Chong Teng. 2024. Reverse multi-choice dialogue commonsense inference with graph-of-thought. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19688–19696.

Li Zheng, Donghong Ji, Fei Li, Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, and Chong Teng. 2025b. Ecqed: emotion-cause quadruple extraction in dialogs. *IEEE Transactions on Audio, Speech and Language Processing*.

Li Zheng, Tengyue Song, Yuzhe Ding, Xiaorui Wu, Fei Li, Dongdong Xie, Jinbo Li, Chong Teng, and Donghong Ji. 2025c. Improving emotion and intent understanding in multimodal conversations with progressive interaction. *IEEE Transactions on Affective Computing*.

Li Zheng, Sihang Wang, Hao Fei, Zuquan Peng, Fei Li, Jianming Fu, Chong Teng, and Donghong Ji. 2025d. Enhancing hyperbole and metaphor detection with their bidirectional dynamic interaction and emotion knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL'25)*.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257.