

# Do Large Language Models Have “Emotion Neurons”? Investigating the Existence and Role

Jaewook Lee<sup>1,2\*</sup> Woojin Lee<sup>1\*</sup> Oh-Woog Kwon<sup>2</sup> Harksoo Kim<sup>1†</sup>

<sup>1</sup>Konkuk University, Republic of Korea

<sup>2</sup>Electronics and Telecommunications Research Institute, Republic of Korea

{benecia428, writerwoody}@gmail.com

ohwoog@etri.re.kr nlpdrkim@konkuk.ac.kr

## Abstract

This study comprehensively explores whether there actually exist “emotion neurons” within large language models (LLMs) that selectively process and express certain emotions, and what functional role they play. Drawing on the representative emotion theory of the six basic emotions, we focus on six core emotions. Using synthetic dialogue data labeled with emotions, we identified sets of neurons that exhibit consistent activation patterns for each emotion. As a result, we confirmed that principal neurons handling emotion information do indeed exist within the model, forming distinct groups for each emotion, and that their distribution varies with model size and architectural depth. We then validated the functional significance of these emotion neurons by analyzing whether the prediction accuracy for a specific emotion significantly decreases when those neurons are artificially removed. We observed that in some emotions, the accuracy drops sharply upon neuron removal, while in others, the model’s performance largely remains intact or even improves, presumably due to overlapping and complementary mechanisms among neurons. Furthermore, by examining how prediction accuracy changes depending on which layer range and at what proportion the emotion neurons are masked, we revealed that emotion information is processed in a multilayered and complex manner within the model.

## 1 Introduction

Emotions play a crucial role in human cognition, behavior, and communication (Plutchik, 1980; Lazarus, 2000). As large language models (LLMs) are increasingly utilized in various application domains, the ability to understand and appropriately respond to human emotions is becoming ever more important. Conventional methods have primarily

assessed LLMs’ emotional capabilities using metrics such as classification accuracy in emotion classification tasks (Zhang et al., 2023) or response generation that reflects specific emotional states (Chen et al., 2023). However, these outcome-based evaluations have been criticized for not directly illuminating *how* emotion information is processed and represented *within* an LLM.

Against this backdrop, the present study systematically investigates both the existence and functional role of neurons (hereafter referred to as “**emotion neurons**”) within LLMs that *selectively* process and represent specific emotions. This concept extends the “sentiment neuron” proposed by Radford et al. (2017), examining whether large-scale models contain sets of neurons that are *actually* highly responsive to certain emotions and, if so, how these neurons function in terms of overlap, distribution across layers, and so on. In particular, this study focuses on the conversational context, where emotion processing is critically important in human–AI interaction, and adopts the six basic emotions (happiness, sadness, anger, disgust, fear, surprise) as proposed by Ekman (1992).

To this end, we leveraged representative LLMs and a synthetic dialogue dataset containing explicit emotion expressions to explore **three key research questions (RQ1, RQ2, RQ3)** in a step-by-step manner. First, **RQ1** asks, “Do emotion neuron groups actually exist within LLMs that are responsible for particular emotions?” To address this, we conducted a detailed analysis of neuron activation patterns in each layer of the model, identifying neuron groups that consistently activate under specific emotions. Next, **RQ2** examines whether these “emotion neurons,” once identified in RQ1, truly serve a functional role by observing how the model’s emotion prediction performance changes when they are manipulated. Specifically, we investigated how *overlaps among emotion neurons* and *complementary relationships between different*

\*Main contributors

†Corresponding author

*emotions* influence performance variations.

Finally, in **RQ3**, we aimed to refine our understanding of *where* emotion neurons are distributed within the model and *how* they integrate emotion information by examining, from multiple angles, how manipulating emotion neurons at different *layer ranges* and at different *ratios* affects the model’s emotion prediction accuracy. Rather than simply treating the entire set of emotion neurons as “present/absent,” this involved exploring changes in the model’s emotion processing that arise from varying both the intensity and the layer-based position of neuron control, thus offering a deeper understanding of how LLMs internally represent emotions.

Our experimental results confirmed that a significant number of neurons related to specific emotions exist within LLMs, and that intentionally manipulating these neurons leads to a marked decline in performance for those emotions. Moreover, the distribution of these neurons varies depending on the scale of the model and the depth of its architecture, and we observed strong overlaps in certain emotions, indicating complementary effects among the emotion neurons. Finally, the analysis focusing on particular layer ranges and manipulation ratios showed that the location and intensity of these emotion neurons have a decisive impact on the model’s emotion prediction capabilities.

## 2 Related Work

With recent rapid advances in LLMs, there has been a notable improvement in performance across a wide range of natural language processing tasks (Achiam et al., 2023; Zhao et al., 2023). These developments have demonstrated *near human-level* language understanding and generation capabilities across diverse application domains, prompting researchers to probe the limitations of LLMs more deeply (Gallegos et al., 2024; Huang et al., 2023). In particular, questions have arisen about whether LLMs can respond in ways analogous to human emotional states and to what extent they possess human-like emotional understanding. This has driven active research into LLMs *in the context of human–AI interaction* (Chang et al., 2024; Xi et al., 2023).

**Evaluation of LLMs’ Emotional Intelligence.** Such interest has naturally led to academic inquiries into how much *Emotional Intelligence* (EI) LLMs possess, an ability crucial for human–AI

interaction. For instance, Wang et al. (2023) established evaluation criteria for EI in various LLMs to quantify each model’s capabilities for emotion recognition and response, thus examining how well they handle emotional cues in actual conversations. However, Paech (2023) pointed out that existing EI measurement methods might overestimate or underestimate the models’ true capabilities. In response, they introduced a new evaluation dataset called *EQ-Bench* to address these issues. Similarly, Sabour et al. (2024) developed an independent benchmark called *EmoBench*, enabling detailed analyses through categorized emotional indices. Meanwhile, Lee et al. (2024) extended these approaches to Vision Large Language Models (VLLMs), conducting an in-depth study of which factors are crucial for emotion recognition when processing a combination of visual and linguistic inputs.

### **Methods for Improving Emotional Recognition.**

As the ability to recognize and appropriately respond to human emotions has emerged as a core capability for conversational AI, a variety of methodological efforts have been made to enhance such abilities (Lei et al., 2023; Liu et al., 2024). For example, it has been reported that applying a *Mixture of Experts (MoE)* architecture—commonly used to improve overall LLM performance—to emotional recognition tasks can yield performance gains by assigning *specialized expert* modules for each emotion category, which then work in a *mutually complementary* manner (Lim and Cheong, 2024). Furthermore, Li et al. (2024) introduced an *Emotional Chain-of-Thought (ECoT)* approach to refine emotional dialogue generation, employing a strategy that systematically breaks down emotional cues in the conversation. In a similar vein, Zhang et al. (2024) proposed a Set-of-Vision (SoV) prompting method to boost the emotion recognition performance of VLLMs.

### **Limited Understanding of Emotion Neurons.**

Until now, most research has focused on enhancing *emotion recognition performance* in LLMs. However, there has been relatively little investigation into the fundamental question of “*At which stage and by which neuron groups is emotion information actually processed within an LLM?*” In this paper, we identify and manipulate “*emotion neurons*” in the model, systematically analyzing whether (1) there truly is a group of neurons that handle particular emotions and (2) how controlling these neurons alters the model’s emotion recognition perfor-

mance.

### 3 Methodology

This section outlines the detailed procedure used to identify emotion neurons within an LLM and analyze their functionality. We first provide an overview of the model architecture that grants access to emotion neurons, followed by a description of the core techniques for detecting and selecting such neurons. All methodological steps described here were applied to the dataset introduced in Appendix B.

#### 3.1 Model Architecture Overview

Modern LLMs generally adopt a Transformer-based *decoder-only* architecture, and in this study, we analyze an *open-source LLM* that follows this design. When an input sentence is processed, it is first tokenized into a sequence of tokens  $t = \langle t_1, t_2, \dots, t_n \rangle$ . Each token is then mapped to an embedding vector  $x_i$  via an embedding matrix  $W_E \in \mathbb{R}^{|V| \times d}$ . The resulting embedding sequence  $x = \langle x_1, x_2, \dots, x_n \rangle$  is fed into the model  $f$ .

The model is composed of  $L$  layers, each consisting of (1) multi-head self-attention (MHA) and (2) a feed-forward network (FFN). As the token embeddings pass through the layers, they are iteratively updated into higher-dimensional contextual representations (i.e., *residual stream states*). After all layers have been processed, the final representations are projected into the vocabulary space via an unembedding matrix  $W_U$ , and a softmax function is applied to produce a probability distribution over the next token.

#### 3.2 Multi-head Self-attention (MHA)

Multi-head self-attention (MHA) enables tokens within an input sequence to *dynamically* reference *different positions*, leading to richer contextual representations. In layer  $l$ , MHA first projects the input representation  $X^{l-1}$  into queries ( $Q$ ), keys ( $K$ ), and values ( $V$ ):

$$Q = X^{l-1}W_Q^{l,h}, K = X^{l-1}W_K^{l,h}, V = X^{l-1}W_V^{l,h}. \quad (1)$$

For token  $i$ , the dot product between its query  $Q_i$  and the key  $K_j$  of every token  $j$  is scaled by  $\sqrt{d_h}$  and passed through a softmax, yielding attention weights  $a_{i,j}^{l,h}$ :

$$a_{i,j}^{l,h} = \text{softmax} \left( \frac{Q_i K_j^\top}{\sqrt{d_h}} \right). \quad (2)$$

These weights indicate the relative importance of token  $j$  for updating token  $i$ . The weighted sum of the values ( $V$ ) produces  $\text{Attn}^{l,h}(X^{l-1})$ . The outputs from all heads are concatenated, transformed via a linear mapping, and then added back to the residual stream, updating the input for the next step  $X^{\text{mid},l}$ .

#### 3.3 Feed-forward Network (FFN)

After the attention process, the FFN applies additional nonlinear transformations to refine token representations. It typically consists of two linear transformations with a nonlinear activation function in between:

$$\text{FFN}^l(X^{\text{mid},l}) = g(X^{\text{mid},l}W_{\text{in}}^l)W_{\text{out}}^l, \quad (3)$$

where  $g(\cdot)$  governs *neuron activation*. Neuron  $u$  is considered activated if

$$n_u^l = \max(0, h_u^l), \quad (4)$$

where  $h_u^l$  is the linear output before activation. In this study, to identify “emotion neurons” that respond to specific emotions, we focus on the activation patterns of individual neurons within the FFN layers.

#### 3.4 Emotion Neuron Selection

Drawing inspiration from the neuron-level analysis method described in Tang et al. (2024), this study identifies neurons within the model that strongly respond to particular emotions by performing the following steps. This approach constitutes the first step in quantitatively addressing **RQ1** regarding the existence of emotion neurons, and the same set of identified neurons is employed in subsequent analyses for **RQ2** and **RQ3**.

**Measuring Activation Frequency by Emotion.** Based on the six basic emotions proposed by Ekman (1992), we prepared dialogue data labeled with six emotions ( $\mathcal{E} = \{\text{anger, disgust, fear, happiness, sadness, surprise}\}$ ). For each FFN layer neuron  $n_u$ , we calculate how often it is activated ( $\max(0, \cdot)$ ) when sentences labeled with a given emotion  $e$  are provided, denoting this count as  $f_{u,e}$ .

**Calculating Emotion-wise Activation Probability.** Let  $T_e$  be the total number of tokens labeled with emotion  $e$ . Then the probability  $P_{u,e}$  that neuron  $n_u$  is activated under emotion  $e$  is

$$P_{u,e} = \frac{f_{u,e}}{T_e}. \quad (5)$$

We then apply  $L1$  normalization across all emotions to obtain an emotion distribution  $\hat{P}_{u,e}$  for each neuron:

$$\hat{P}_{u,e} = \frac{P_{u,e}}{\sum_{e' \in \mathcal{E}} P_{u,e'}}. \quad (6)$$

**Using Entropy for Emotion Neuron Determination.** The more  $\hat{P}_{u,e}$  is skewed toward a specific emotion, the more likely it is that neuron  $u$  specializes in that emotion. We measure this via entropy  $H_u$ :

$$H_u = - \sum_{e \in \mathcal{E}} \hat{P}_{u,e} \log \hat{P}_{u,e}. \quad (7)$$

A lower  $H_u$  means  $\hat{P}_{u,e}$  is concentrated on a particular emotion. Therefore, we select as *emotion neurons* the top 1% of neurons with the lowest  $H_u$ .

## 4 Results

### 4.1 RQ1: Do emotion neuron groups actually exist within LLMs that are responsible for particular emotions?

In this section, we investigate whether emotion neurons within an LLM truly exist to handle particular emotions. We conducted an emotion-wise neuron distribution analysis on two models, *Llama-3.1-8B-Instruct* and *Llama-3.1-70B-Instruct*. If emotion neurons do exist, we should be able to identify neuron groups in the model that display *consistent activation patterns* linked to specific emotions (e.g., *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*). To verify this, we quantitatively measured the extent to which activation is concentrated on each emotion per layer and then visualized the distribution across layers in Figure 1.

**Existence of emotion neurons and differences in neuron distribution for each emotion.** Our analysis showed that neurons with a strong correlation to particular emotions exist in both models above a certain threshold. In *Llama-3.1-8B-Instruct*, the emotion with the largest total number of identified neurons was *anger* (1,882 neurons), followed by *fear* (1,629), *disgust* (1,598), *sadness* (1,570), *happiness* (1,320), and *surprise* (1,070). This indicates that neuron groups for particular emotional categories are distinct and suggests a mechanism of emotional representation in LLMs akin to the “sentiment neuron” proposed by Radford et al. (2017).

In *Llama-3.1-70B-Instruct*, the absolute number of emotion neurons increased overall, yet *surprise*

(4,976) remained the smallest distribution. Interestingly, in the 70B model, *sadness* (8,314) surpassed *anger* (7,910) as the emotion with the most neurons, confirming that the relative distribution can shift as the model size grows. This suggests that both training procedures and the growth in parameters can influence the model’s internal emotional representation structure.

**Layer-wise distribution patterns of emotion neurons.** Examining how emotion neurons are distributed across layers revealed that both models exhibit a “distribution curve” in neuron activation, rather than a uniform pattern throughout the layers.

- *Llama-3.1-8B-Instruct*: Emotion neuron activation is relatively low in the initial layers, increases sharply in the middle layers, and then diminishes again toward the later layers, with a slight uptick near the final layer. This suggests that the mid-layer regions may play a central role in integrating contextual and emotional information.
- *Llama-3.1-70B-Instruct*: Emotion neurons exhibit relatively high activation from the earliest layers, span the middle layers, and gradually diminish in the upper layers. Compared to the 8B model, having deeper architecture and more parameters may allow emotional information to be distributed and processed from the earliest stages, eventually being integrated into higher-dimensional linguistic contexts.

**Summary of RQ1 results.** From this analysis addressing **RQ1**, we confirm that neurons can indeed be identified for each emotion in both models. This finding suggests that emotional processing may not be random but is instead concentrated in specific layers or neuron groups. Furthermore, the fact that the number and distribution of emotion neurons can vary with model size indicates that changes in the parameter count directly influence how emotions are represented within the model.

### 4.2 RQ2: What are the functional effects of manipulating emotion neurons on the model’s emotion recognition performance?

In **RQ1**, we established that there are, in fact, sets of neurons within the LLM that are responsible for processing specific emotions and that their distribution patterns differ by emotion. In this section,

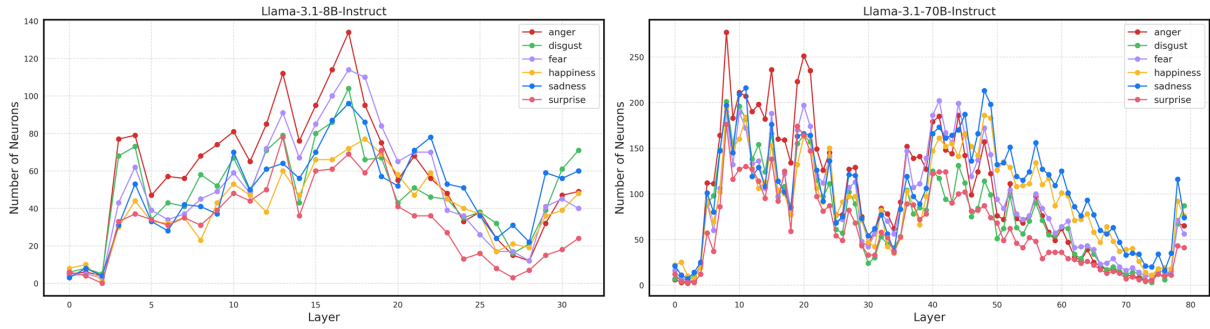


Figure 1: Comparison of the layer-wise distribution of emotion neurons for each emotion in Llama-3.1-8B-Instruct (left) and Llama-3.1-70B-Instruct (right).

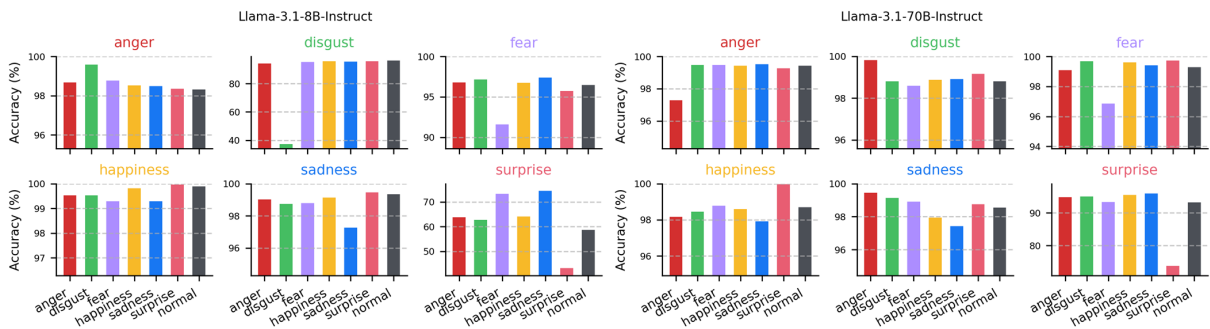


Figure 2: Each subplot shows the classification accuracy for the target emotions (anger, disgust, fear, happiness, sadness, surprise). The x-axis compares the normal (unmodified) condition against the masking of neurons for each specific emotion, illustrating how prediction performance changes when the neuron group for a particular emotion is masked.

we examine **RQ2**. By *manipulating* the previously identified neurons, we observe how the model’s accuracy in classifying the corresponding emotion changes. If masking a certain set of emotion neurons substantially degrades the model’s ability to predict that emotion, it indicates that the neuron set plays a *functionally* important role in processing it. Figure 2 shows the changes in prediction accuracy after masking.

**Effects of emotion neuron manipulation and performance degradation.** Looking first at the results from *Llama-3.1-8B-Instruct*, the *disgust* emotion stands out. In the normal (unmodified) state, the model’s accuracy for predicting *disgust* was 96.11%, but this dropped significantly to 37.22% upon masking *disgust* emotion neurons. This clearly demonstrates that even in a smaller model (8B), neurons for *disgust* indeed exist, and once removed, the model becomes unable to properly recognize it. Similarly, for the *surprise* label in *Llama-3.1-70B-Instruct*, the accuracy was 93.17% under normal conditions but declined dramatically to 73.83% (a 19% decrease) when *surprise* neurons

were masked. This indicates not only that *surprise* neurons were identified in RQ1, but also that removing them causes a sharp decline in the model’s ability to predict that emotion. In other words, these neurons appear to play a *direct* role in classifying *surprise*.

**Differences in emotional dependence and overlapping, complementary mechanisms.** Interestingly, not every emotion shows a strong reliance on such clearly defined sets of emotion neurons. For instance, in the 8B model, masking *happiness* neurons yields almost no change in the accuracy of the *happiness* label, and for *anger*, the prediction accuracy even shows a slight increase after masking. This suggests that, for certain emotional categories, when one group of neurons is removed, other neurons compensate, thereby preventing any performance drop. Similarly, in the 70B model, when *disgust* neurons are masked, the accuracy for *disgust* remains nearly the same as the normal state, implying no noticeable performance decline. This suggests that “disgust emotion” may be represented with overlap among other emotion neurons

or that sufficient alternative neurons exist within the model to fulfill a similar function when these neurons are removed.

**Overlap among emotion neurons and functional substitutability.** Further discussion of neuron overlap is provided in Appendix C, where Figure 14 visualizes the *Overlap Ratio* of emotion neurons in both *Llama-3.1-8B-Instruct* and *Llama-3.1-70B-Instruct*. A concise summary is that there can be considerable overlap between neuron sets for different emotions, but a high degree of overlap does not necessarily guarantee functional substitutability.

When interpreting the present experimental results, note that “no observable performance drop after masking” should not be hastily interpreted as the *absence* of specific emotion neurons. For example, in the 70B model, although accuracy for the *disgust* label remains high after masking *disgust* neurons, it would be premature to conclude that the model has *no* neurons for disgust. It is entirely possible that overlapping, complementary mechanisms among various neurons are at work, allowing other neurons to compensate for the lost functions.

Conversely, a high overlap ratio does not inherently assure “functional substitutability.” For instance, in *Llama-3.1-8B-Instruct*, there is a relatively high overlap ratio of 48% between *anger* and *disgust* neurons, yet if *disgust* neurons are removed, we see a sharp decline in accuracy from 96.11% to 37.22%. This is a prime example of how “**high overlap**  $\neq$  **strong substitutability**.” Even if a single neuron *responds* to multiple emotions, it does not necessarily mean its contribution is equally significant for all of them. Consequently, removing neurons that are functionally critical to a particular emotion can substantially compromise the model’s performance.

**Summary of RQ2 results.** In summary, while masking emotion neurons leads to a marked decline in prediction accuracy for certain emotional categories (*surprise*, *disgust*, etc.), some emotions (*happiness*, *anger*, etc.) show virtually no performance loss—and may even improve—indicating a more complex pattern. These findings have several implications:

- **Confirmation of the reality of emotion neurons:** For certain emotions (e.g., *surprise*, *disgust*), removing the corresponding neurons directly impairs performance, suggesting that

these neurons play an essential role in processing those emotions.

- **Overlap among emotion neurons:** In some cases (e.g., *happiness* in the 8B model, *disgust* in the 70B model), performance does not drop significantly after neuron removal, implying that various neuron sets operate via overlapping, complementary effects.
- **No necessary equivalence between high overlap and functional replacement:** As seen in the *anger–disgust* example, even if the overlap ratio is high, other emotion neurons might not fully replace the primary function of the removed neurons, indicating that “high overlap” does not necessarily mean “high substitutability.”

#### 4.3 RQ3: How do the *masking ratios* and *layer-based manipulations* of emotion neurons affect emotion prediction performance?

In RQ1 and RQ2, we found that *emotion neurons* exist in both *Llama-3.1-8B-Instruct* and *Llama-3.1-70B-Instruct* and that *selectively manipulating* these neurons leads to changes in emotion prediction accuracy. In this section, we focus primarily on the larger *Llama-3.1-70B-Instruct* model to investigate how (1) varying the ratio of manipulated emotion neurons and (2) masking neuron sets in different layer ranges influence emotion prediction. Figures 3 and 4 illustrate how the model’s emotion prediction performance changes when neurons associated with particular emotions are masked, relative to the normal (unmodified) condition.

#### Changes in prediction performance according to varying *masking ratios* of emotion neurons.

As shown in Figure 3, for certain emotions, step-wise increases in the degree to which emotion neurons are suppressed (masked) produce a notably “stepped” decline in accuracy for that emotion. For instance, masking only 1% of anger neurons resulted in about a  $-2\%$  drop in accuracy, but raising that to 4–5% suppression amplified the drop to around  $-4\% \sim -5\%$ . This finding aligns with RQ2, suggesting that the *functional significance* of emotion neurons becomes more evident at higher *masking ratios*.

In particular, *surprise* shows a steep decline of more than  $-20\%$  in accuracy from masking

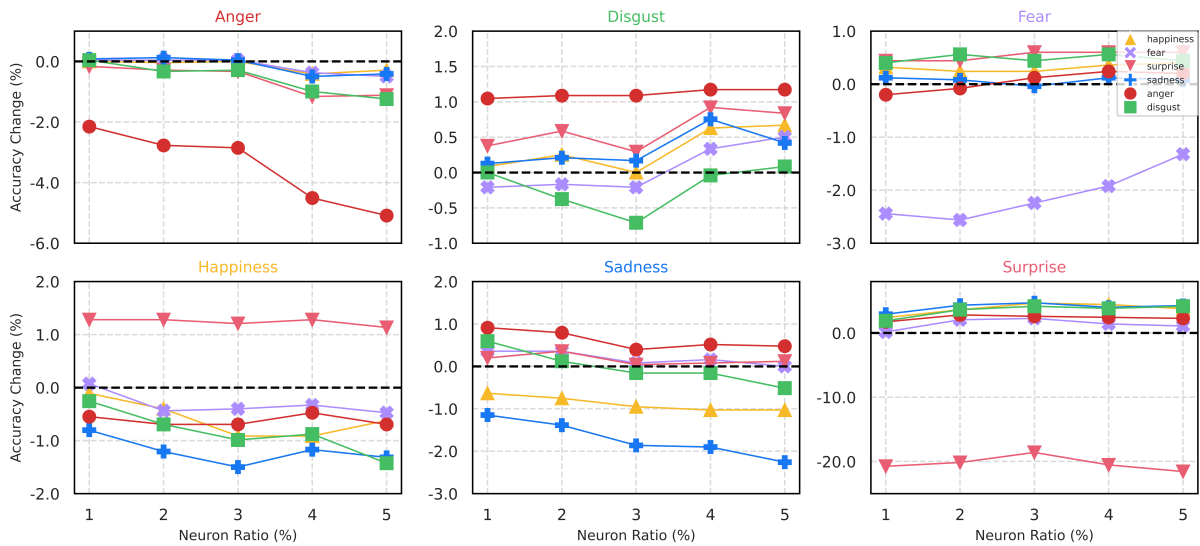


Figure 3: Visualization of how classification accuracy changes for the corresponding emotion and for other emotion labels when the masking ratio of a particular emotion neuron set is gradually increased from 1% to 5%.

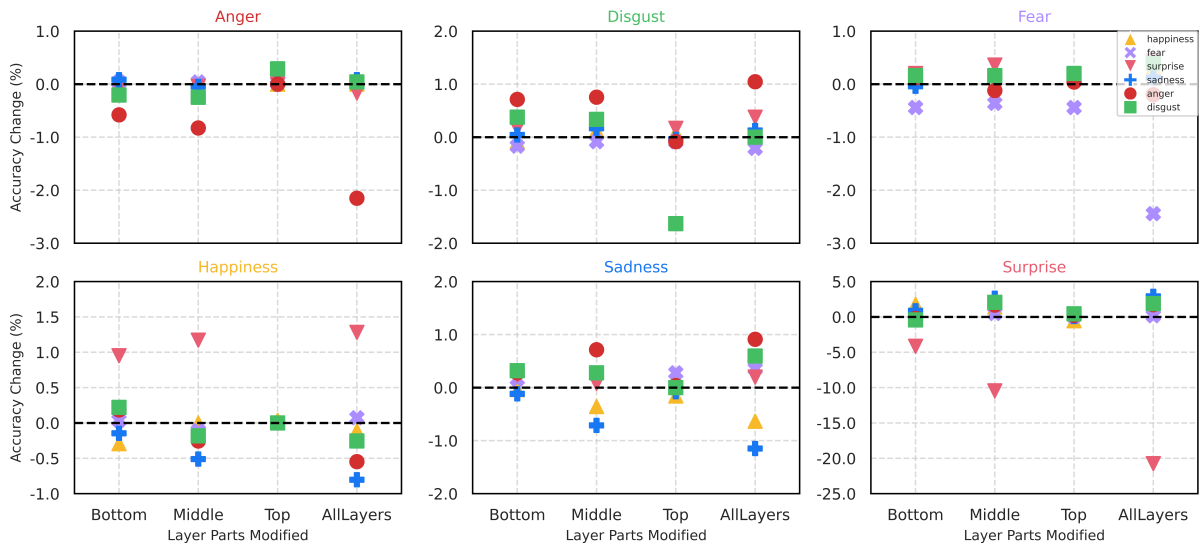


Figure 4: Comparison of how emotion prediction performance changes when emotion neurons are selectively masked at Bottom, Middle, Top, or AllLayers.

just 1% of its neurons, with a continued large decrease ( $-21\% \sim -22\%$ ) at higher levels of suppression. This indicates that recognizing surprise is highly dependent on a specific set of neurons. By contrast, disgust accuracy fluctuates only mildly at high suppression levels, and in some cases, the cross-emotion accuracy even *improves*. Such “overlap/substitution effects” are consistent with the overlapping, complementary mechanism discussed in **RQ2**, implying that some emotions (e.g., disgust) can be partly managed by other neuron sets when the primary ones are removed.

ing one particular emotion resulted in *improved* accuracy for another emotion label often confused with it—for example, masking surprise neurons slightly raised accuracy for happiness. This can be interpreted as indicating that certain emotion neurons operate competitively or share overlapping representations, such that suppressing one set makes it easier for the model to distinguish another. Overall, the results for varying masking ratios reinforce the idea that “*emotion prediction performance changes in complex ways depending on how strongly neurons are suppressed,*” and they underscore that *overlap/substitution mechanisms*

Additionally, cases were found where suppress-

differ by emotion.

**Effects of layer-based masking on emotion prediction accuracy.** Examining Figure 4 shows that the degree of change in prediction accuracy can vary substantially depending on *which layers* are manipulated, even if the same emotion neurons are masked. For instance, if *surprise* neurons are removed only in the Bottom layers, the accuracy decline is moderate (−4.15%). However, when the Middle layers are manipulated, accuracy drops dramatically (−10.46%). Further, masking the same emotion neurons across AllLayers yields an even more substantial reduction (−20.75%), suggesting that *surprise* emotion is handled by neurons distributed across multiple layers.

Conversely, *disgust* shows a pattern in which accuracy may even *increase* when the Bottom layers are suppressed, and in AllLayers masking, performance often remains nearly unchanged (close to 0%). However, if we concentrate on the Top layers, accuracy decreases more distinctly (−1.63%), indicating that *disgust* may undergo final processing in the upper layers. This suggests that *disgust* neurons could be forming overlapping, complementary relationships with other emotional signals in the lower and middle layers, making it difficult to interpret the full processing of disgust by focusing on just one layer range.

These differences in the primary distribution and function of emotion neurons, depending on emotion and layer, further confirm from a more fine-grained perspective the points raised in **RQ1** and **RQ2** regarding “heterogeneous neuron distribution patterns” and “changes in functional capacity when these neurons are manipulated.” For example, *fear* neurons show a relatively small drop in accuracy when masked only in the Bottom or Top layers, but accuracy declines considerably (−2.44%) when they are removed from AllLayers. One possible interpretation is that the neuron group processing *fear* is spread across various layers, so partial removal still leaves enough neurons to fulfill the function, whereas masking them in all layers sharply impairs performance.

**Summary of RQ3 results.** Combining the experiments on masking ratios and layer-based manipulation, we can draw the following points:

- **Masking ratio perspective:** For some specific emotions, a stepwise increase in suppression of emotion neurons leads to a progressive

decline in prediction performance for those emotions, re-confirming the substantive importance of those neurons. However, certain emotions allow for partial or minimal decline, or even an improvement in cross-emotion performance, indicating that an *overlapping and substitutive mechanism* may be at work.

- **Layer-based perspective:** While emotion neurons are dispersed across multiple layers, some emotions (e.g., disgust) show a marked drop in prediction performance primarily when Top layers are masked, implying the possibility of “key layers” for each emotion. Additionally, masking AllLayers typically leads to the greatest overall accuracy decline, suggesting that emotion neurons do not operate in isolation within a single layer but function in an overlapping fashion throughout a multi-layer structure.

## 5 Conclusion

This study systematically explored whether neurons exist within LLMs that process specific emotions (RQ1), how the model’s emotion recognition function changes when these neurons are selectively removed (RQ2), and how emotion neurons are distributed across layers, as well as how changing masking ratios and layers affects prediction accuracy (RQ3). The experimental findings revealed significant sets of neurons for major emotions, with certain emotions showing a dramatic decrease in prediction accuracy when their neurons were removed, indicating these neurons serve a functionally important role. At the same time, we observed that overlapping and substitutive mechanisms operate among emotion neurons, so that for some emotions, model performance did not significantly drop or even improved, demonstrating that emotional representation is organized by mutually complementary structures rather than a simple one-to-one mapping. Moreover, stepwise masking or selectively manipulating certain layer ranges resulted in varying effects on emotion prediction, suggesting a complex, multi-layer mechanism in which emotional information is dispersed, integrated, and processed.

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (RS-2025-00553041,



Enhancement of Rational and Emotional Intelligence of Large Language Models for Implementing Dependable Conversational Agents, 50%) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. RS-2023-00216011, Development of Artificial Complex Intelligence for Conceptually Understanding and Inferring like Human, 30%) and IITP grant funded by the Korea Government (MSIT) (No. RS-2024-00338140, Development of learning and utilization technology to reflect sustainability of generative language models and up-to-dateness over time, 20%).

## Limitations

By identifying emotion neurons in LLMs and examining their functions in a systematic manner, this study contributes fresh insights into how LLMs represent emotions internally. Nevertheless, it is important to acknowledge the following limitations, which point to intriguing directions for future research:

First, we focused on the six basic emotions (happiness, sadness, anger, disgust, fear, surprise). While this approach allows us to clearly delineate which emotion neurons to investigate, real human emotions are far more nuanced and can be continuous. A variety of emotional states exist, such as *jealousy*, *embarrassment*, and *relief*, and it remains possible that corresponding neurons also exist within LLMs. Future work could extend and refine these analyses by employing more granular emotion models or multidimensional emotion representations.

Second, this study centers on emotion recognition in text-based dialogues; however, real-world scenarios often involve various multimodal signals, such as *vocal tone*, *facial expression*, and *gestures*. How such multimodal inputs might affect the activation of emotion neurons in an LLM or VLLM remains unexplored. Further research is needed to examine how emotion neurons manifest and can be manipulated in environments where multimodal data are presented in parallel.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Richard S Lazarus. 2000. How emotions influence performance in competitive sports. *The sport psychologist*, 14(3):229–252.

Jaewook Lee, Yeajin Jang, Hongjin Kim, Woojin Lee, and Harksoo Kim. 2024. Analyzing key factors influencing emotion prediction performance of vllms in conversational contexts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5801–5816.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Dongjun Lim and Yun-Gyung Cheong. 2024. Integrating plutchik’s theory with mixture of experts for enhancing emotion classification. In *Proceedings of the*

2024 *Conference on Empirical Methods in Natural Language Processing*, pages 857–867.

- 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.
- Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.
- Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1.
- Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.
- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.
- 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*.
- Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.
- Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. 2024. Visual prompting in llms for enhancing emotion recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4484–4499.
- Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

## A Experimental Setup and Implementation Details

**Models.** For the experiments in this study, we employed two models: *Llama-3.1-8B-Instruct* and *Llama-3.1-70B-Instruct* (Dubey et al., 2024). Both models were pretrained and instruction-tuned and are known to have strong language comprehension and reasoning abilities in conversational contexts. To compare and analyze differences in the distribution and functionality of emotion neurons based on model size (i.e., parameter count), both models were tested under identical experimental conditions.

**Prompt Configuration.** To evaluate emotion prediction performance and explore emotion neurons, the same *System Prompt* and *User Prompt* were used for both models. Figure 5 illustrates an example of the prompts. The placeholder {prompt} indicates where the actual dialogue text for which we want to predict emotion is inserted. The user provides specific dialogue examples at that position, and the model outputs a single emotion label (from among the six categories). Through this prompt structure, we directed the model to select and output *only one emotion* in response to the explicit instruction.

## B Synthetic Data Generation for Emotion Neuron Exploration

A large-scale conversational dataset containing diverse emotions is essential for conducting the emotion neuron exploration experiments proposed in this study. However, publicly available conversation corpora often include multiple emotions in a single dialogue, making it difficult to secure enough examples for any single emotion. This can hamper the process of identifying and analyzing neurons responsive to specific emotions. To address this issue, we compiled a separate synthetic dataset in which each dialogue is constrained to contain exactly one emotion. The data generation followed three main steps: **(1) Topic Augmentation**, **(2) Generating Emotion-Infused Dialogues**, and **(3) Filtering the Generated Dialogues**. The overall process is summarized in Figure 6, and the prompts used in each step are shown in Figure 7.

### B.1 Topic Augmentation

To maximize the diversity of conversation topics, we started with a base pool of 315 topics. We then

used the GPT-4o model to repeatedly generate new alternative words or phrases for each topic, eventually building a total pool of 5,040 topics.

A sample prompt for this step was: “Please generate a one- or two-word alternative to the term ‘{topic}’”, designed to produce simple, non-duplicative words or phrases semantically related to {topic}. Sample transformations, such as “New York” → “Los Angeles” and “Apple Company” → “Microsoft”, were provided so the model could—via a few-shot learning approach—generate suitable alternatives. Additionally, constraints such as “Avoid terms that add descriptive phrases or explanations” were imposed to prevent introducing unnecessary detail.

Through this procedure, the initial 315 topics doubled to 630 after the first transformation, then to 1,260 upon the second, ultimately reaching 5,040 topics. This expanded topic pool served as the foundation for generating diverse emotion-infused dialogues in the next step.

### B.2 Generating Emotion-Infused Dialogues

From each of the 5,040 topics obtained in the topic augmentation step, we automatically generated dialogues infused with one of the six basic emotions—happiness, sadness, anger, disgust, fear, surprise—as proposed by Ekman (1992), underpinning our research focus on identifying emotion neurons.

For reasons related to model usage costs and accessibility, we used the *gemini-1.5-flash* model for dialogue generation. For every topic, we generated 10 dialogues per emotion, thus 60 per topic, totaling  $5,040 \times 6 \times 10 = 302,400$  dialogues.

A representative prompt for this generation process was: “As a creative writer, craft a natural and coherent dialogue between two characters about the topic ‘{topic}’”, with an additional instruction: “The conversation should vividly convey the emotion of ‘{emotion}’ throughout the interaction.” We also imposed the constraint: “Use the following format strictly without any additional text or explanations”, so that the content was produced strictly in the form “A: [First character’s utterance] / B: [Second character’s response]”, facilitating consistency for subsequent processing.

No exact duplicates were detected in the experiment. Representative samples of generated dia-

logues are provided in Figures 8 through 13, listed in the order of happiness, sadness, anger, disgust, fear, and surprise.

### B.3 Filtering the Generated Dialogues

Finally, to confirm that the generated dialogues indeed matched both the target emotion and topic, we carried out a *filtering* procedure. Specifically, we used three different proprietary (private) models—*gpt-4o-mini*, *gemini-1.5-flash*, and *claude-3-haiku*—to independently predict the emotion present in each generated dialogue.

Thus, four labels were assigned to each dialogue: (1) the single emotion label specified during dialogue generation, and (2) three predicted labels from the three models. Only those dialogues for which at least three of the four labels matched the originally assigned emotion were retained in the final dataset. This conservative criterion served as a minimum standard to ensure that the dialogue truly embodied the given emotion. Consequently, 8,685 dialogues were filtered out. After filtering, a total of  $302,400 - 8,685 = 293,715$  dialogues remained; their distribution by emotion is shown in Table 1.

In addition, 95% of the final dataset was used for emotion neuron exploration and model training, while the remaining 5% was reserved for evaluating changes in prediction performance following emotion neuron masking.

## C Analysis of Emotion Neuron Overlap

In this section, we examine the overlap patterns among the sets of identified emotion neurons in each model, based on the heatmap shown in Figure 14. Specifically, we investigate the extent to which the neuron sets corresponding to six emotions (anger, disgust, fear, happiness, sadness, surprise) overlap in *Llama-3.1-8B-Instruct* and *Llama-3.1-70B-Instruct*.

**Degree of overlap among emotion neurons and symmetrical structure.** In Figure 14, the rows and columns follow the same emotion order (anger, disgust, fear, happiness, sadness, surprise). Because the row (*reference emotion*) and column (*comparison emotion*) form a symmetrical relationship, the overlap ratio between *anger* and *disgust* is identical to that between *disgust* and *anger*, for instance.

**Overlap patterns in specific emotion pairs and their implications.** A salient finding is that in

both models, the *anger–disgust* pair exhibits a relatively high overlap ratio (around 48% in the 8B model and 43% in the 70B model), suggesting that anger and disgust may share somewhat similar negative emotional signals (e.g., pronounced negativity). Likewise, both models display a relatively high overlap ratio between *fear* and *sadness*, indicating that these two emotions may share certain negative or passive features.

By contrast, happiness shows consistently low overlap with other emotions. For instance, in the 8B model, the overlap ratios between *happiness* and negative emotions (*anger*, *disgust*) are only around 4–8%, and in the 70B model, the overlap between *anger* and *happiness* is about 3%. This suggests that *happiness* neurons may remain comparatively *independent* from negative emotion neurons.

**Comparison of overlap based on model size (8B vs. 70B).** For most emotion pairs, the 8B model shows higher overall overlap ratios than the 70B model. This implies that a smaller model with fewer parameters may reuse or merge multiple emotion representations in a *more compact* way, whereas the 70B model likely has a more *distributed and specialized* neuron space for each emotion. This aligns with **RQ1**, which observed that the distribution of emotion neurons becomes more varied as model size increases.

**Relationship between overlap and functional substitutability.** In the neuron masking experiment from **RQ2**, for some emotions (*disgust*, *surprise*, etc.), prediction accuracy deteriorated severely after neuron removal, indicating that these neurons are *essential* for those emotions. However, this does not imply that “low overlap immediately results in performance loss, whereas high overlap guarantees complete substitutability.” Indeed, for *anger–disgust* in the 8B model, even though the overlap ratio is high, removing *disgust* neurons still caused a significant performance drop. This demonstrates that even if neurons are shared, their functional contributions to each emotion may differ.

**Partial overlap and the structure of emotion neuron representation.** Overall, these results indicate that an LLM’s emotion neurons are not necessarily composed of entirely distinct sets but can exhibit considerable *mutual overlap*. This does not contradict **RQ1**’s finding that “**each emotion’s neuron set exists**”; rather, it clarifies that multi-

ple emotions may share and utilize some signals within *one model*. Notably, certain negative emotion pairs (e.g., *anger–disgust*, *fear–sadness*) show high overlap ratios, whereas positive emotions like *happiness* typically exhibit lower overlap. This suggests that similar emotions may be processed by comparable neuron patterns in the model, while more distinct emotions remain relatively separable.

Still, high overlap does not necessarily guarantee functional substitutability, as **RQ2** showed that some neurons can be partially replaced by others, whereas removing certain neuron sets causes a severe performance deficit. Thus, even if neurons are shared, the manner in which each emotion is weighted or activated can differ substantially.

## D Additional Misclassification Analysis under Emotion Neuron Masking

Here, we quantitatively examine which emotion labels the input samples of a certain emotion shift to when those neurons are masked. Figure 15 displays a heatmap of these shifts. For each model (*Llama-3.1-8B-Instruct*, *Llama-3.1-70B-Instruct*), the rows represent the *masked emotion*, while the columns represent the *erroneously predicted emotion*, maintaining the same ordering for both dimensions. Each cell indicates the extent to which misclassified samples (for the masked emotion) cluster into a particular erroneous emotion label. Note that these are distributions *within the failed cases only*, excluding any “correctly predicted” samples. Below, we discuss key confusion patterns observed in specific emotion pairs and their possible explanations.


**Confusion between anger and disgust.** In both models, when anger neurons are masked, there is a very high tendency for misclassifications to switch to disgust; conversely, when disgust neurons are masked, errors predominantly shift to anger. This suggests these two emotions may share a similar negative affect in the model’s internal representation. Once neuron masking loosens the boundary between them, the model’s decision is more likely to default to the related emotion.

**Unidirectional substitutions between happiness and surprise.** When happiness neurons are masked, most misclassifications become surprise, and when surprise is masked, errors often shift to happiness—indicating a strong bidirectional relationship. This could be interpreted as reflecting


that both emotions embody a relatively heightened positive affect, making it easier for the model to rely on the remaining similar emotional representation after one set is removed. It also reinforces the view that happiness and surprise may be closely connected in the model, contrasting with more negative emotions.

**Diverse misclassification pathways after masking fear.** When fear neurons are masked, the smaller model tends to confuse it mainly with anger or sadness, whereas the larger model exhibits a broader distribution of errors, including anger, disgust, sadness, and surprise. This suggests that as model size increases, neuron sets for each emotion become more widely distributed across layers and time steps, leading to a more varied range of possible misclassification outcomes when a specific set of neurons is removed.

**Overlap and differentiation revealed by misclassification rates.** In both models, once neuron sets are masked, errors mostly shift to other emotions that share similar features (e.g., *anger–disgust*, *fear–sadness*, *happiness–surprise*). However, across the larger positive–negative divide, cross-confusion is rare, indicating that the internal representation of contrasting affective categories remains fairly well separated. This aligns with the neuron overlap analysis discussed earlier, suggesting that certain emotion pairs are tightly connected within the model, whereas mechanisms governing opposing emotions may have fewer interaction points.



## Used Prompt



### Emotion Prediction

**System Prompt:**  
 You are a helpful assistant that predicts emotions from dialogues. Your task is to predict the emotion conveyed in a given dialogue.

You must strictly follow these guidelines:

1. Predict one of the following emotions only: [anger, disgust, fear, happiness, sadness, surprise].
2. Output only the predicted emotion in the requested format, without any additional text, explanations, or comments.

**User Prompt:**  
 Given the following dialogue, predict the emotion that is most clearly conveyed. The dialogue is structured as follows:  
 {prompt}  
 Your output should be in the following format:  
 [One of anger, disgust, fear, happiness, sadness, surprise]

Figure 5: Example configuration of the System Prompt and User Prompt for exploring emotion neurons and predicting emotions.

Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Total
<b>Count</b>	49,015	49,232	50,166	55,054	51,549	38,709	293,725
<b>Percentage (%)</b>	16.68	16.76	17.08	18.74	17.54	13.18	100.00

Table 1: Emotion counts and percentages after LLM evaluation.

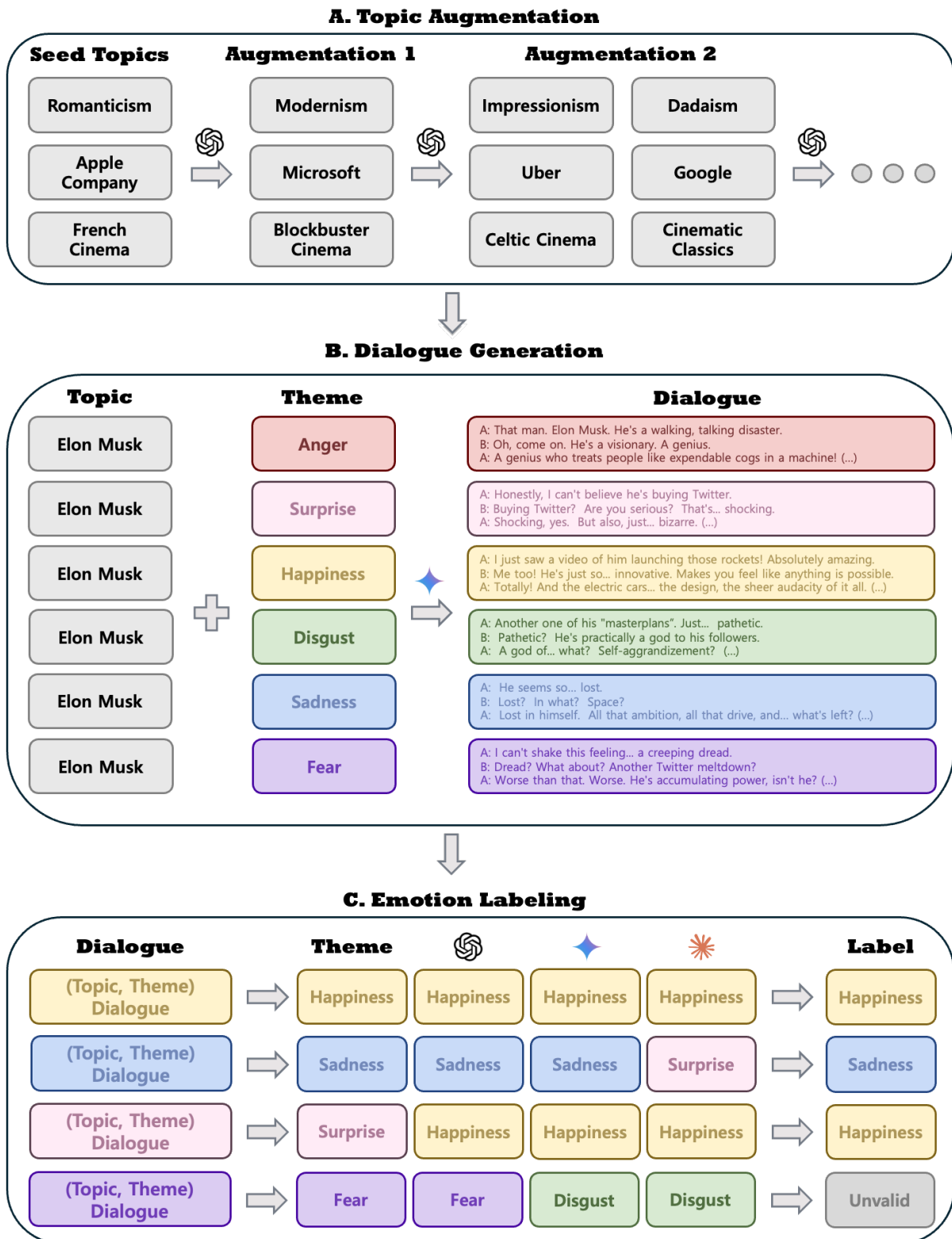


Figure 6: Entire process for generating the large-scale dialogue dataset used in emotion neuron exploration experiments.



## Prompt Design for Synthetic Data Generation



### Dialogue Topic Generation with FITS (gpt-4o-mini)

Please generate a one- or two-word alternative to the term "{topic}" that is semantically related and could replace it in similar contexts. Requirements: 1. The new term should be relevant to "{topic}" and evoke similar associations without being identical in form or concept. 2. Avoid terms that add descriptive phrases or explanations, such as "The History of" or "Overview of." Instead, aim for concise, commonly recognized nouns or phrases. 3. Ensure the alternative is distinct from, and not included in, the following existing list: {' , 'join(existing\_topics)} Example Transformations: - "New York" → "Los Angeles" - "Apple Company" → "Microsoft" - "European Union" → "NATO" - "Piano" → "Violin" Provide only the new term without any additional explanations or annotations.



### Synthetic Dialogue Generation (gemini-1.5-flash-8b)

As a creative writer, craft a natural and coherent dialogue between two characters about the topic "{topic}". The conversation should vividly convey the emotion of "{emotion}" throughout the interaction. Use the following format strictly without any additional text or explanations: A: [First character's utterance] B: [Second character's response] A: [First character's reply] ... Ensure that the dialogue captures the essence of "{emotion}" and stays focused on the topic "{topic}". Do not include any narration, descriptions, or emotion labels. Provide only the dialogue in the specified format.



### Synthetic Data Labeling - initial (gemini-1.5-flash, claude-3-Haiku, gpt-4o-mini)

You are a sentiment analysis expert. Analyze the following conversation and determine the primary emotion that best represents the overall sentiment expressed throughout the dialogue. Select only from the following emotions: {' , 'join(emotions)}. Conversation: {dialogue['dialogue']} Please respond with only the emotion name from the provided list. Do not include any additional text, formatting, or explanations. .



### Synthetic Data Labeling - unknown (gemini-1.5-flash, claude-3-Haiku, gpt-4o-mini)

You are a sentiment analysis expert. Analyze the following conversation and determine the primary emotion that best represents the overall sentiment expressed throughout the dialogue. **\*\*Important: You must select only one emotion from the following list and respond with exactly one word in lowercase. Do not use any other emotion words or additional text, formatting, explanations, or comments.\*\*** List of emotions: {' , 'join(emotions)}. Conversation: {dialogue['dialogue']} Please respond with only the emotion name from the provided list. Do not include any additional text, formatting, or explanations.

Figure 7: Prompts used to create the dialogue data.



Emotion: Happiness  
Topic: Harvard University



I can't believe we actually got in! Stanford!



Seriously? Oh my goodness, I'm so happy. It's unbelievable.



The campus looks absolutely stunning in the pictures. I'm imagining myself walking through the quad, surrounded by all those brilliant minds.



Me too! And the libraries. I've heard they're amazing. Just thinking about it makes me smile.



I know! And the research opportunities. The possibilities are endless.



Totally! I'm picturing myself in a lab, surrounded by all the equipment, making groundbreaking discoveries.



Exactly! It's just... perfect.



Perfect! I can't wait to experience it all.



Me neither. I'm already picturing myself in my dorm room, surrounded by my new friends.



Oh, that's so exciting. I can already feel the energy of the student body.



The excitement is overwhelming, but in a wonderful way. I'm so excited to be a part of it all.



Me too. I'm bursting with happiness. This is going to be incredible.



It truly is. This is the start of a beautiful future.



Yes, a fantastic future. I'm so, so happy.

Figure 8: A dialogue generated based on the happiness emotion.

Emotion: Sadness  
Topic: Movie Star



Remember how radiant she was? Like the sun breaking through clouds.



Yeah, a fragile beauty. A fleeting light.



Now... just a whisper of a memory. Dust motes dancing in the fading afternoon.



The paparazzi's harsh glare, the endless parade of faces... it all chipped away at her.



She was always so... vulnerable. So exposed.



And the roles, they became a cage. Each one a tiny, gilded prison.



They never truly understood her. Never saw beyond the shimmering facade.



The whispers followed her, like shadows. They wouldn't let her be.



Now, just a name on a list of departed stars. A faded poster in a forgotten shop.



The applause, the cheers, the adoration... all gone silent. Just echoes in the empty theatre.



Such a beautiful, tragic story. A life lived too fast.



And the price... so steep. A heart aching in the spotlight.



It's a heartbreak, isn't it? A silent, shimmering heartbreak.

Figure 9: A dialogue generated based on the sadness emotion.

Emotion: Anger  
Topic: Cooking



This stew is inedible! Burnt to a crisp, like a forgotten offering to a vengeful god.



It's just a little...dark. A touch of...character.



Character? It's a culinary crime scene! You've ruined the entire pot!



I followed the recipe. Precisely.



Precisely wrong! You didn't even taste it, did you?



I did. It tasted...rich.



Rich? Rich like burnt sugar and disappointment!



Maybe a little...experimentation is needed.



Experimentation? You're a culinary terrorist!



I'm trying to perfect the art.



The art of ruining perfectly good vegetables!



They'll be fine. We can always add more seasoning.



More seasoning? You think adding salt to a charcoal briquette will make it palatable?



Perhaps a little more...depth.



Depth? The depth of despair! I'm never eating anything you cook again!



Oh, I'm sure you'll be hungry soon.



You think I'll be hungry? I'll be starving for something edible!



I'll just...go get some takeout.



Go get takeout? After this disaster? You're a disgrace to the kitchen!

Figure 10: A dialogue generated based on the anger emotion.

Emotion: Disgust  
Topic: Simplification



This whole "simplification" thing is nauseating.



Nauseating? It's practical. Streamlined.



Streamlined? Like a butchered souffle! Everything's reduced to... what? Beige? Bland?



Efficient. Focused.



Focused on what? Losing the nuance? The heart? The soul?



The soul? We're talking about spreadsheets, not symphonies.



And symphonies are somehow less important? This relentless pursuit of the least common denominator... it's repellent.



It's progress.



Progress towards... what? A sterile wasteland? A world where everything tastes the same?



It frees us from... complexity.



From complexity? From the very thing that makes us... human? This reductionism is... a horror.



You're being overly dramatic.



Dramatic? It's a betrayal! A betrayal of beauty, a betrayal of... everything.



Beauty? Beauty is subjective. Simplification is objective.



Objective? It's a... a... a... disgustingly, aggressively, objectively... wrong thing to do.



You're just clinging to the past.



No. I'm clinging to what makes life worth living. And this... this... simplification... it's sucking the life out of everything.

Figure 11: A dialogue generated based on the disgust emotion.

Emotion: Fear  
Topic: Skin Cancer



I saw a new mole on my arm yesterday. It's...different.



Different how?



It's...it's a little darker, and it feels...rough.



How long has it been there?



Just a couple of weeks. But it's...it's growing.



You should really get it checked. A doctor.



I know, I know. But what if it's bad?



What if it's not?



But what if it is? What if it's melanoma? I keep seeing those pictures online.



Don't look at pictures. Focus on getting it looked at.



It's hard not to. They're everywhere. Scary pictures.



I know. It's easy to get overwhelmed. But think about it this way. Early detection is key.



But what if it's too late? What if I've already...what if I've already missed something?



You haven't. You're going to get it checked. Now.



I wish I could just...erase it. Make it go away.



You can't, but you can take care of it. You'll be okay. Promise.



Promise? What if you're wrong?



I'm not wrong. We'll figure this out. Together.



I'm scared, you know? Really scared.



I know you are. But we'll get through this. Just...take a breath.

Figure 12: A dialogue generated based on the fear emotion.

Emotion: Surprise  
Topic: SpaceX



I can't believe SpaceX is launching another mission to Mars.



Really? Wow. I thought they were focused on the Starship program for a while.



I know, right? It's just... so sudden. I was expecting a lot more testing.



Totally. This is... unexpected. Are they even ready?



Well, they've been so efficient with the Falcon 9 launches, but... Mars? That's a whole different ballgame.



You're right. I mean, the sheer scale of the project. It's astounding.



And the cost! I heard the ticket prices are... well, astronomical.



Astronomical? Seriously? I thought they were trying to make space travel more accessible.



Maybe they're just trying to... shock us.



Shock us? That's... a bold strategy. I'm flabbergasted.



I know. This is all so surprising. I just... I wasn't expecting this.



Me neither. This is... quite a surprise package.

Figure 13: A dialogue generated based on the surprise emotion.

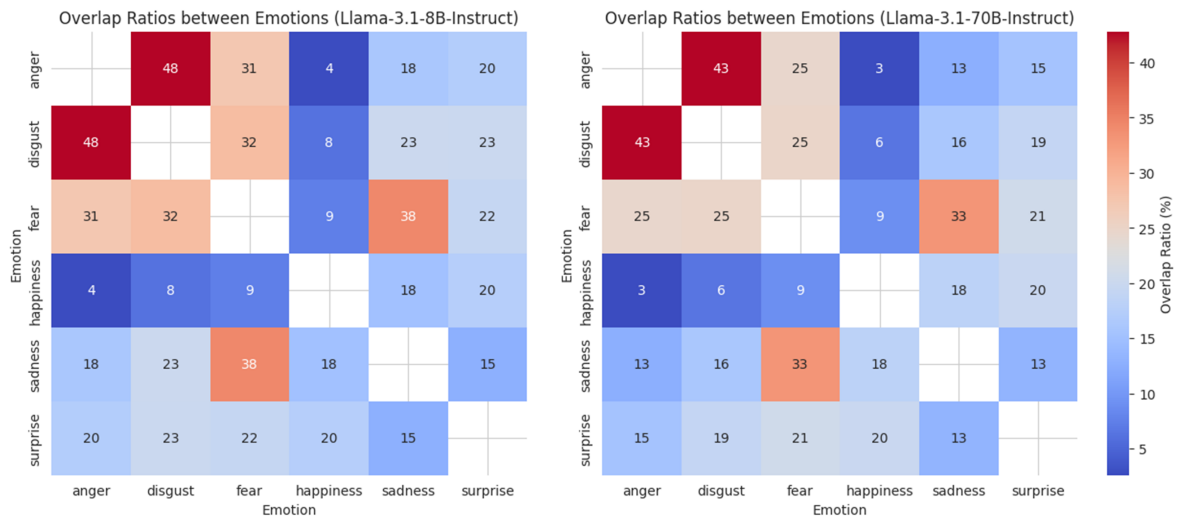


Figure 14: Heatmap of overlap ratios among sets of emotion neurons.

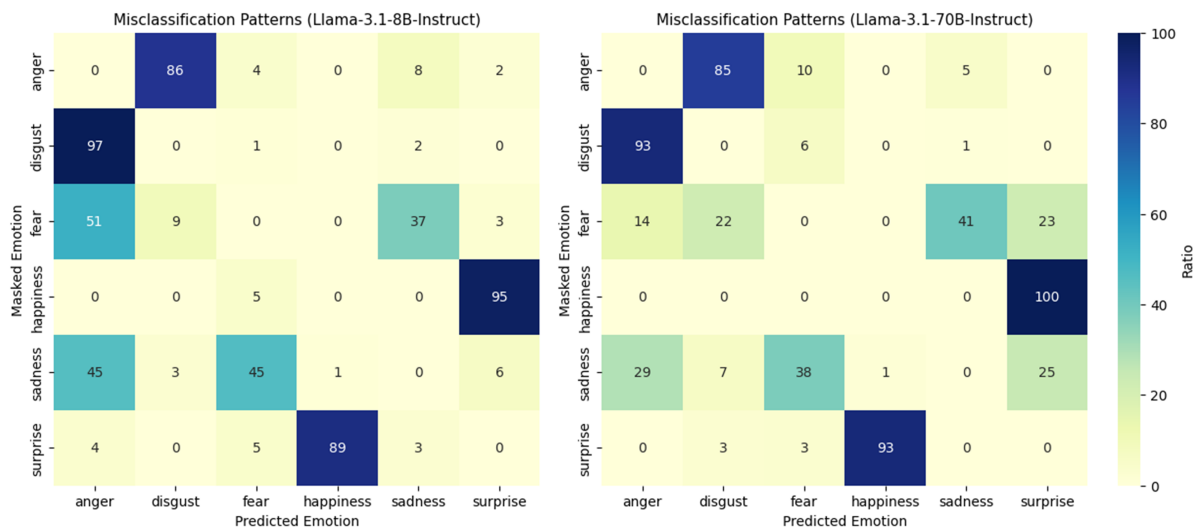


Figure 15: Heatmap of misclassification patterns after emotion neuron masking.