A Unified View on Emotion Representation in Large Language Models

Aishwarya Maheswaran
IIT Hyderabad, India
ai21resch11002@iith.ac.in

Maunendra Sankar Desarkar
IIT Hyderabad, India
maunendra@cse.iith.ac.in

Abstract

Interest in leveraging Large Language Models (LLMs) for emotional support systems motivates the need to understand how these models comprehend and represent emotions internally. While recent works show the presence of emotion concepts in the hidden state representations, it's unclear if the model has a robust representation that is consistent across different datasets. In this paper, we present a unified view to understand emotion representation in LLMs, experimenting with diverse datasets and prompts. We then evaluate the reasoning ability of the models on a complex emotion identification task. We find that LLMs have a common emotion representation in the later layers of the model, and the vectors capturing the direction of emotions extracted from these representations can be interchanged among datasets with minimal impact on performance. Our analysis of reasoning with Chain of Thought (CoT) prompting shows the limits of emotion comprehension. Therefore, despite LLMs implicitly having emotion representations, they are not equally skilled at reasoning with them in complex scenarios. This motivates the need for further research to find new approaches[1].

1 Introduction

Fine-tuning transformer models, such as RoBERTa and BERT, has demonstrated high levels of accuracy on popular sentiment and emotion classification benchmarks like (Maas et al., 2011), (Socher et al., 2013), and (Saravia et al., 2018). With the advancement of Large Language Models (LLMs), there was a natural shift in leveraging these models for the same task, and comparable or better performances were achievable without the need for explicit training (Zhang et al., 2024). This motivates the need to examine these models and understand how they perform this task. Techniques from Mechanistic Interpretability (MI) (Bereska and Gavves, 2024) and Representational Engineering (RE) (Zou et al., 2025) help us understand the inner working of models by looking at their representations. We then try to understand the emotion reasoning ability of LLMs, using a challenging dataset (Sabour et al., 2024) involving multiple emotions and perspectives. Recent works have attempted to identify the specific range of layers where LLMs represent emotions. (Tak et al., 2025) showed emotion concepts to be present in the intermediate layers of the model while (Palma et al., 2025) showed it to be in the early layers. We hypothesize that this discrepancy is due to the difference in the datasets and specific prompts used. In this work, we leverage interpretability techniques to present a unified view on where LLMs represent emotions by experimenting with multiple diverse datasets and prompts on two model families. We then evaluate if the existence of such representations helps the reasoning ability (via CoT) in LLMs.

The experiments reveal the following key insights: (i) The Type of instruction prompt plays a key role in determining the layers where emotion concepts are represented. (ii) Compressing the model's hidden state representations into emotion-specific direction vectors (using the technique from (Zou et al., 2025)) exhibits high cosine similarity across datasets, indicating a consistent representation of underlying emotion concepts. (iii) The vectors representing emotion obtained from one dataset can be utilized across datasets for the pairwise emotion classification task with minimal impact on subsequent prediction accuracies, which highlights the *consistent functional nature* of the representations. (iv) While CoT predicted answers mostly overlap with predictions without CoT, we observe the LLM is more prone to change its answer with CoT when its initial entropy over the emotion possible choices is high. This reveals the role of the model's uncertainty in being influenced by reasoning traces. The design of the experiments towards having an over-

[1] Code available at GitHub repo

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3589–3610
March 24-29, 2026 ©2026 Association for Computational Linguistics

all understanding of where and how the LLMs can comprehend the emotions, and the associated findings are the main contributions of the work. The findings further emphasize the requirement of in-depth processing and reasoning of the emotion concepts for handling complex scenarios that require emotional understanding.

2 Related Works

2.1 Emotion representation in LLMs

Residual Stream - The hidden state representations from the intermediate layers are shown to encode representations of emotion concepts detectable with linear classifiers. Sentiment is shown to have a linear direction representation (Tigges et al., 2024). (Palma et al., 2025) showed such representation for emotions in the early layers and sentiment in the later layers in LLMs. For a different dataset (Tak et al., 2025), it was shown that emotions are detectable in the intermediate layers. This motivates our first experiment, which applies supervised probing to diverse datasets and prompts. (Zou et al., 2025) showed these representations can be distilled into emotion reading vectors (which encode the space the model used to represent that emotion internally). We utilize this method in our second set of experiments to show that the emotion reading vectors are consistent across datasets.

Attention and Feed Forward layers - (Tak et al., 2025) claim, the representations from attention layers are responsible for shaping the emotion representation detected in the Residual stream. They compare with the representations from Feed Forward Network (FFN) and show that the accuracies are similar to that of the residual stream.

Neurons - These are intermediate representations in the FFN layer. A neuron is considered activated (relevant to the task) if it has a positive value after the application of non linearity function in the FFN layer. Neurons that are active for Sentiment and Emotion classification tasks were identified in the early to middle layers of LLMs (Song et al., 2024). The location and density of these neurons seem to be emotion-specific (Lee et al., 2025).

2.2 Techniques used

Probing - It is a technique of training a classifier on the hidden representation obtained from the model. It can be done in a supervised setting (linear classifiers, SVM, Decision trees, etc.) or with an unsupervised setup (PCA/K-means). Probing has been used in several recent works such as (Palma et al., 2025), (Tak et al., 2025), (Tigges et al., 2024) and (Zou et al., 2025). These are the methods we use for our experiments.

Activation Patching - It is a method of replacing a particular activation/representation with a corresponding representation obtained for a different input. The impact on the output is indicative of the role of the representation and the location where it was patched. This was used in (Tak et al., 2025) and (Tigges et al., 2024).

Entropy and Gradient methods - It calculates the entropy/gradient of neurons for different labeled datasets to identify the task or concept-specific neurons. It is used in (Lee et al., 2025) and (Song et al., 2024).

Although several works have used interpretability techniques to understand emotion identification capabilities of LLMs, there are contradictions in the findings across the works (Tak et al., 2025), (Palma et al., 2025). We opine that these contradictions arise due to the use of different prompts and datasets having different difficulty levels. Our intent is to analyze the setup using prompts with different expressivity levels across datasets with different difficulty levels – to get an overall understanding of the ability of LLMs in identifying emotions at their representation level. We additionally try to assess the reasoning ability of such models with an emotion comprehension task.

3 Assessing Emotion Comprehension Abilities of LLMs

We perform several approaches to understand the emotion comprehension abilities of LLMs. The experiments are structured as follows:

1. We perform *Supervised Probing* on different intermediate representations (Residual, Attention, MLP) for two different model families on four different emotion-labeled datasets, to detect the emotion present in the input. We experiment with three instruction prompt variations to observe its impact on the emotion representation.

2. We use *Unsupervised Probing* to obtain a representation of emotion concept space in the form of emotion reading vectors. This helps us to compare the emotion representations across datasets.

3. We analyze the impact of Chain of Thought (CoT) prompting on emotion classification for a complex dataset - EmoBench (Sabour et al., 2024). This helps us understand if LLMs are capable of

Feature	Activation Patching	Neuron Level Analysis
Conceptual Constraints	Ideally, activation patching requires a pair of similar prompts (in format/context) with variation only at the point of interest. For the emotion identification task, such a setup is challenging, especially when comparing across various datasets. Patching success rises only when the patching spans multiple consecutive layers. This implies that emotion information is distributed across layers, which diminishes the utility of patching localized activations.	Neurons often encode multiple concepts, making it difficult to guarantee the identified neurons are purely emotion-specific. Ex: Higher overlap of identified emotion-specific neurons across datasets with P1 compared to P0 indicates the common neurons across emotions are potential instruction-related neurons. It can also encode multiple emotions, representing a more general concept. Ex: Early layers share neurons identified for anger, joy, and sadness.
Prompt Issues	Prior works approach of patching the last token representation, which corresponds to the emotion, is unsuitable for the P0 prompt (which intends to observe naturally elicited emotion representation) in this work, as the resulting last token representation does not necessarily contain the emotion representation.	Unstable across prompts. The emotion categories represented by individual neurons can vary with variation in prompts for the same dataset and model. Final layer shows an increase in identified neurons when P1 (inst prompt) is used, compared to no prompt (P0).
Expt. Results	Experiments with prompt P2 showed low patching success with the MLP and Attention layer representations (even with span=5). High success only in Residual Stream.	Across datasets, although each layer contains roughly 100-200 neurons identified as emotion specific, the number of shared neurons is in the range of 10 per layer.
Takeaway	**Focus on Residual Stream:** To change the model's mind, we essentially have to rewrite the residual stream in the later layers, where the model's focus is on output generation. Hence, we chose to emphasize results from the residual stream in this work.	**Prefer Probing:** (a) Low overlap of emotion-related neurons across datasets, (b) Overlap of emotions within a single neuron, and (c) The difference in emotions associated with a single neuron when the prompt is changed, lead us to prefer probing-based approaches in this work.

Table 1: Comparison of Challenges with Activation Patching and Neuron Level Analysis

generating reasoning that is accurate and helps improve it's emotion classification accuracy.

Experiments with other standard interpretability techniques like activation patching and neuron level analysis had some constraints that made probing more suitable method for this work. We elaborate the challenges with these approaches in Table 1. The results of activation patching are presented in Appendix E.3 and results of emotion neurons identification in Appendix E.4.

3.1 Datasets

To get a unified understanding, we use five different emotion identification datasets. The datasets are: Twitter Emotions Dataset (Saravia et al., 2018), RECCON (Daily Dialog split) Dataset (Poria et al., 2021), an LLM-generated Dataset introduced in (Zou et al., 2025), Crowd-eVent Dataset (Troiano et al., 2023). For CoT analysis we use EmoBench Dataset (Sabour et al., 2024). The description and dataset size details are mentioned in Appendix A.

3.2 Models

We run supervised probing experiments on instruct version of the following LLMs: Llama-3.2-Instruct 1B and 3B, Llama-3.1-Instruct 8B (Touvron et al., 2024), and Gemma-2-it 2B and 9B (Gemma Team, 2025). This helps us see how emotion comprehension abilities for classification change with model size and families. For the unsupervised probing experiments and analysis of the CoT experiments, we focus on Llama 8B and Gemma 9B. We also compare the CoT accuracies with those of Llama 1B and Gemma 2B to examine the effect of CoT on relatively smaller models.

4 Experiments

4.1 Supervised Probing

For every model and dataset, we use prompts P0: no instruction, P1: instruction to identify the emotion, P2: instruction to identify the emotion along with few shot examples (refer to Appendix B) to obtain the layerwise hidden state representations of the last token from the model. These prompt variations allow us to compare the implicit and explicit representations of the model, as shown in (Maheswaran et al., 2025). The motivation for using the last token in described in Appendix C. We then train a logistic regression classifier[2] ,similar to (Tak et al., 2025). We used the cuml.linear_model library for this purpose. By comparing the results across datasets and different instruction prompts, we are able to present a unified understanding of where emotion representations are observed in models. We perform this experiment on 5 models of various sizes and different families. We show the test classification accuracy heatmap plots for these models with the hidden state representations in Figures 1, 2, 9, 10, and 11.

4.2 Length matched prompt ablation

To assess whether differences in probe accuracy between P1 and P2 stem from prompt length, we introduced a new prompt, P1-long, which preserves P1's intent but is length-matched to P2 (differing by only one token across both model families). We observe that P2 continues to yield higher probe accuracy

[2] Classifier parameters: {C=1, max_iter=1000, tol=1e-4, penalty='l2', solver='qn', fit_intercept=True}. For each layer, a separate classifier is trained.

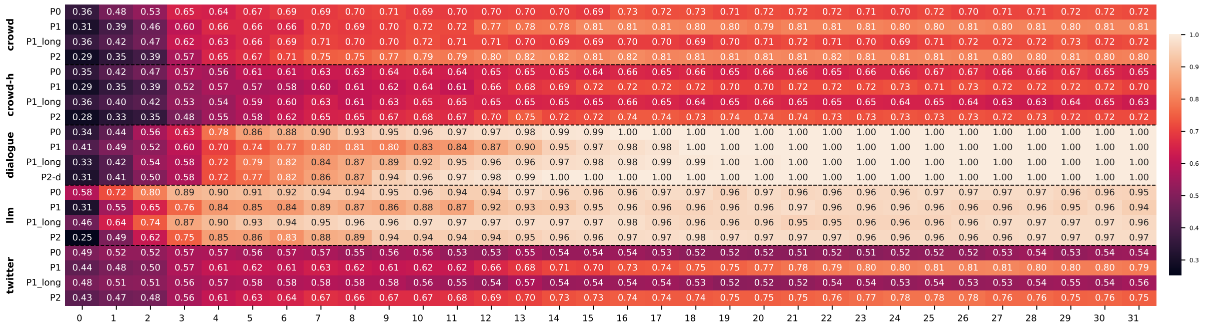


Figure 1: Probing Llama-3.1-8B 6-emotion results

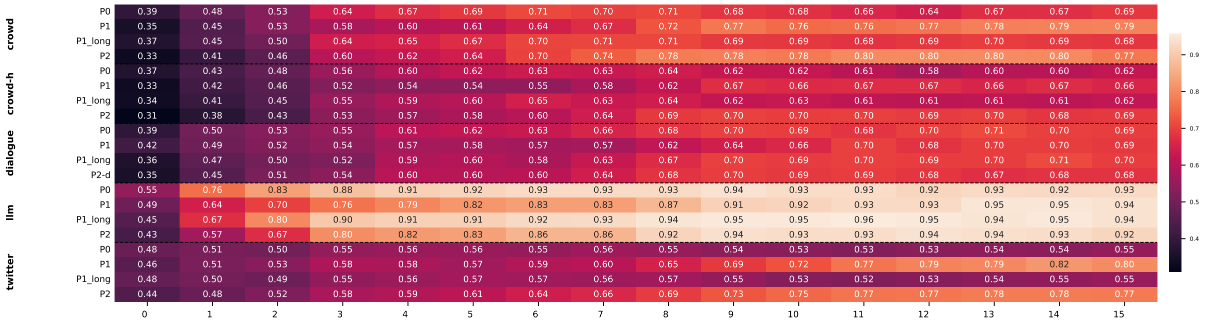


Figure 2: Probing Llama-3.1-1B 6-emotion results

than P1-long in the later layers for the Crowd-eVent and Twitter datasets. Increasing length alone (as in P1-long) negatively affected probe accuracy—an effect not seen with P2. This indicates that token length alone does not explain the differences in layerwise probe performance, disproving the hypothesis that the accuracy gap between P1 and P2 is solely attributable to prompt length.

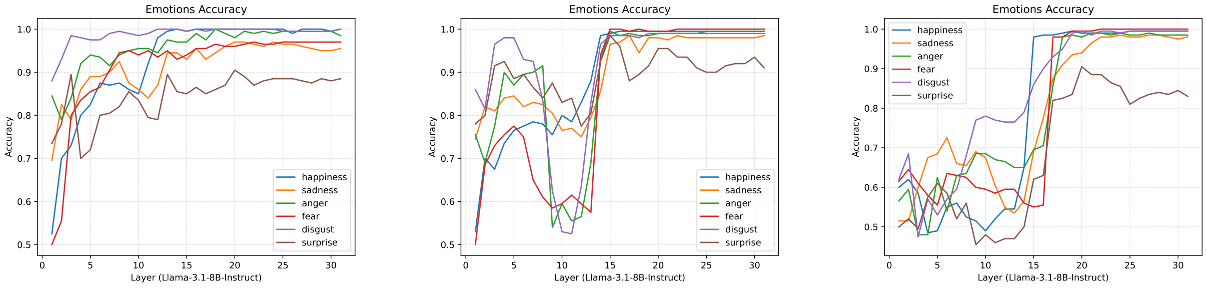
4.3 Unsupervised Probing

We adapt the method of extracting reading vectors described in (Zou et al., 2025) to extract vectors that encode a specific emotion. The data is preprocessed into pairs of the form: $D^e_{train} = (x^+_i, x^-_i)^N_{i=1}$ where e is an emotion of interest, x^+_i is a sample expressing emotion e and x^-_i expresses any other emotion. In our experiments, we set N (the number of samples per emotion) to at most 200, based on data availability. D^e_{test} is prepared similarly, with a different pairing of the data samples. rv_e, the representation of e, is obtained using D^e_{train} following (Zou et al., 2025). Detailed steps of this method is presented in Appendix D due to space limitations.

First, we calculate the reading vectors for one dataset, varying the prompts (P3- Explicitly Mentions the emotion that needs to be focused on in the scenario , P4- Asks to infer the emotion from the scenario, P5- Similar to P4 with the addition of a Speaker label to simulate a dialogue setting. Detailed prompts are given in Appendix B) to understand the impact of the prompt on the representation space. We then calculate these vectors for all the datasets and compare them to measure cosine similarity, shown in Figures 4 and 5. These vectors are used to obtain the scalar projections of the last token representations from the data. For an instance $(x^+_i, x^-_i) \in D^e_{test}$, the emotion prediction is marked as correct iff $proj(x^+_i, rv_e) > proj(x^-_i, rv_e)$. The corresponding results are shown in Table 2 and Figure 3.

4.4 CoT prompting

We evaluate the ability of LLMs in determining emotions (beyond the six emotions (Ekman et al., 1987) tested in the previous experiments) using the EmoBench dataset (Sabour et al., 2024). The dataset is structured with a scenario, subject, and six possible emotion labels as choices. We prompt different LLMs to determine the appropriate emotion label using prompts P6 (direct prompt that asks the model to identify the emotion) and P7 (P6 with think step by step to trigger CoT), (refer Appendix B). We use a temperature of 0.1 to reduce variability when obtaining the generations and set max_new_tokens to 425. Comparison with temperature 0 is provided in Appendix E.1. We evaluate the resulting generations with the follow-



(a) Prompt with explicit emotion (P3)

(b) Prompt without emotion (P4)

(c) Prompt in dialogue setup (P5)

Figure 3: Effect of instruction prompts on accuracy of emotion reading vectors (prominent in early layers)

ing metrics: **a) Accuracy** - Based on the ground truth, **b) Entropy score** - it is calculated over the probability values of the option indexes. Let $\hat{P}$ contain the probabilities assigned to the six possible option tokens (which are among the tokens in the vocabulary) by the model with the prompt. We apply the softmax function to $\hat{P}$ to convert it into a probability vector P and calculate the entropy using $E(P) = -\sum_{i=0}^{5} p_i.log_2(p_i)$. **c) Top2-diff** - is the absolute value of the difference between the top two probabilities in P, a high value indicates the model is confident in its answer, while a low value indicates confusion. **d) Jensen Shannon Divergence** (JSD) (Lin, 1991) - We use it to measure the difference in the probability distributions (P_{direct} and P_{CoT}) obtained with the direct and CoT prompts respectively. A low value indicates the two probability distributions are similar, and a high value indicates it has changed. Evaluating the generations with these metrics provided insights into how CoT influences the model in its final answer prediction. To test if having the ground truth cause or emotion label as part of the input can direct the model towards the right answer and generate an accurate CoT trace, we add a 'hint' as part of the prompt. We vary the type of hint to be either the true cause explaining the emotion in the scenario or the actual emotion expressed. We vary the position of the hint to see the effect of position (refer to prompts P8, P9 having the hint in early and late positions in Appendix B). Table 3 reports the accuracies for these cases.

5 Observations and Results

5.1 Supervised Probing

First, we compare the impact of variation in the instruction prompt within a dataset. Consider the Figures 1, 2 showing the layerwise accuracies for the logistic regression classifier (probe). Prompt P0, which has no explicit instruction to generate the emotion, still has reasonable (e.g. $> 60\%$) accuracy for some datasets. This indicates that: (1) Emotions are implicitly activated in the model representations. This effect is pronounced in the LLM dataset designed to invoke emotions, which also gets the highest probe classification accuracies across datasets. (2) Prompt P2 which includes few shot examples, helps the model (ex: Twitter dataset) increase its final accuracy, at the cost of lower accuracy in the earlier layers compared to other prompts. This may be due to the model spending more time in comprehending the prompt in the early layers. It shows that the LLMs implicitly represent emotion. The extent and depth to which it's represented might be influenced by the prompt. Prompts explicitly asking to identify emotion encourage the persistence of the emotion concepts until the last layer the model as its useful for next token prediction (NTP). This is the key insight to explain the discrepancy in the findings of (Tak et al., 2025) and (Palma et al., 2025) regarding the layer position where high emotion classification accuracy was observed. The instruction prompt, or lack thereof, determines the function of the layers in the model. *In cases where the emotion is the target, the layers focus on maintaining and enhancing the emotion representation, as it is relevant for predicting the next token. Otherwise, the emotion concepts are still present in the representations, but they can be replaced with other representations relevant for NTP in subsequent layers.*

Across datasets, we observe that emotion accuracies can be high in the early or intermediate layers, with the later layers consistently achieving high accuracies, especially for prompts P1 and P2 in all datasets. (3) The clarity with which emotions are expressed in the input seems to play a key role in determining the layer from which it achieves high

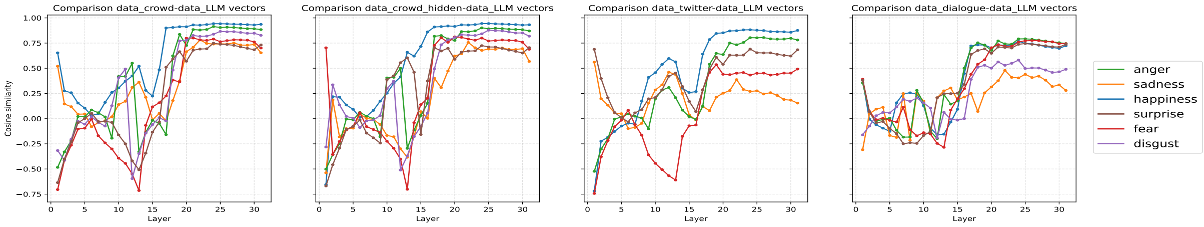


Figure 4: Cosine similarity of reading vectors from different datasets - Llama-3.1-8B-Instruct

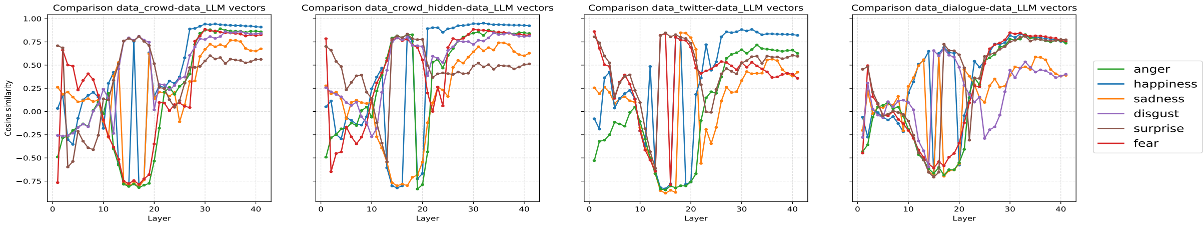


Figure 5: Cosine similarity of reading vectors from different datasets - Gemma-2-9B-it

classification accuracy. The LLM dataset achieves the highest accuracy for all models, as it was carefully created to evoke a particular emotion strongly. Compared to other datasets, where the model acts as an observer trying to infer the expressed emotion based on the context. (4) With the dialogue dataset, the larger model representations are capable of linearly representing emotion from the early layers. In contrast, smaller models from the same family struggle with it. This might be because in a dialog, the individuals can maintain different mental states and there is a frequent switchover between them. Hence, processing the information and identifying the required signals to determine the target emotion requires more work and can be aided by having more number of parameters in the model.

While we focus on representations from the residual stream for the analysis (as noted in Table 1), we present the supervised probing results with representations from the MLP layer and Attention layers in Appendix H.

5.2 Unsupervised Probing

From the supervised probing experiments, it is clear that the model representations encode emotion concepts. Using the reading vectors approach, we distill the representations to obtain emotion reading vectors that encode the direction of a particular emotion for a particular layer. Figure 3 shows the pairwise accuracy obtained by the resulting reading vectors for the LLM dataset for three different input prompts (P3, P4, P5). This confirms our observation from the supervised probing experiment regarding the influence of the instruction prompt in the early layers of the model. Particularly, P3, which explicitly mentions the target emotion, leads the layers to progressively add the emotion concept, increasing accuracy across layers. P4, on the other hand, which does not explicitly mention the emotion, has a drop in the pairwise accuracy after an initial rise in the early layers. However, this drop is quickly rectified, resulting in even higher pairwise accuracies for the individual emotions compared to P3. P5 has the input data in a dialogue format, which is represented differently in the early layers. This clearly shows that while emotions can be represented in early layers, the representation is noisy due to the influence of the prompt. This is an interesting observation because the hidden representations from which the emotion reading vectors are obtained are constructed by taking the difference of contrastive pairs, which should ideally remove the effect of the instruction prompt, as it is common. However, the significant difference in the early layers suggests that the instruction representation still has a lasting impact. The representations in the later layers are consistent and show high pairwise accuracy for most emotions.

Comparing across datasets with a fixed prompt (P4), we observe high pairwise classification accuracy shown in Table 2. To compare if the vectors themselves are similar, we calculate the cosine similarity of the layer-wise emotion reading vectors from different datasets with the vectors from the LLM dataset. Figures 4 and 5 show high similarity, especially for emotions like happiness and anger across datasets. There is low similarity in the early layers with the vectors for the Gemma model, even having opposite directions. But from the intermediate to the late layers, the representations become similar. Since the prompts used to obtain the representations were similar in this case, the difference could be due to dataset differences.

Next, we test if the emotion vectors can be interchanged within datasets. Specifically, we utilize the emotion reading vectors obtained from the LLM dataset, which have demonstrated high pairwise accuracy scores on other datasets. The resulting accuracy values are shown in Table 2. Our results show that *the model has learned a consistent underlying strategy for representing emotion that is similar across datasets (based on the cosine similarity scores) and functionally equivalent for the pairwise classification task.*

5.3 Utility of emotion reading vectors for classification

We find that the emotion vectors obtained by unsupervised probing can act as internal emotion translators, projecting the hidden state representations onto six emotion-specific directions. The magnitude of the projection serves as an indicator of the emotion distribution, with higher projection values indicating a strong representation of that particular emotion, whereas uniformly low or similar projections across emotions indicate a mixed emotional state. **These observations suggest a potential use of these vectors as a lightweight diagnostic tool for emotion analysis**.

5.4 CoT prompting

To understand the impact of additional forward passes provided by CoT on the emotion classification ability, we run experiments comparing Instruct versions of Llama3.2-1B, Llama-3.1-8B, Gemma2-2B, and 9B. Table 3 shows the accuracies (the final selected option index) obtained for direct and CoT prompting. Since we observed significant missing predictions for Llama-1B, we relax the accuracy calculation to include even the final selected option text (if present), wherever the exact index was missing. This is reported in the Acc. (CoT-Aug) column. Considering the highlighted cases, we observe that only Llama-3.1-8B was able to improve its initial accuracy score with CoT. The other models show relatively minor improvements, with Gemma-2-1B experiencing a slight decrease. The agreement between CoT and the direct column shows that Gemma models do not change their predicted answers significantly with CoT, and the improvement gained with CoT is negated by cases where CoT misled the model. Llama models, on the other hand, share relatively fewer predictions between the direct and CoT cases, indicating the model used the CoT to actually change its predictions. We share some examples of the dataset along with the CoT generations by the models in Table 7.

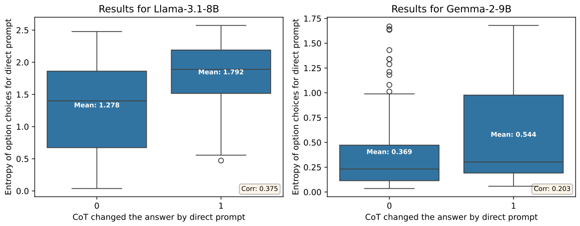


Figure 6: Boxplot of Entropy scores grouped by whether CoT changed its answer

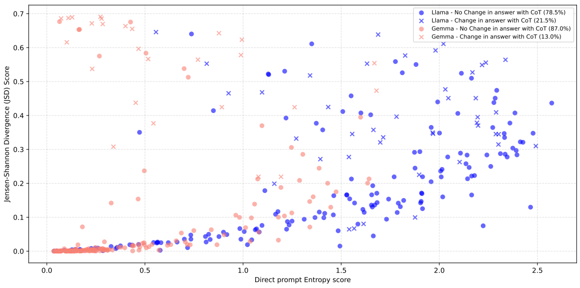


Figure 7: JSD score vs Entropy from direct prompt

- **Effect of initial Entropy over the role of CoT**: To analyse the reason CoT was more computationally active in one setting, we look at the entropy scores over the probability of only the possible option tokens in Table 4. 1) CoT reduced the entropy for both models, indicating that even if the predicted option is the same, the model is more confident with CoT. 2) The entropy of Llama-8B with the direct prompt is higher compared to Gemma. This suggests that the model may be confused and uses CoT to select the appropriate answer. The boxplot in Figure 6 visually confirms that for cases where CoT was able to change the prediction, the initial entropy (with the direct prompt) was higher. Figure 7 plots the individual samples based on their initial entropy score with the direct prompt against the JSD scores. The shape of the marker indicates if CoT had changed the answer. We observe that when CoT changes the answer for Gemma, it does so with a high JSD score. For Llama, which has high initial entropy, CoT can serve as a tiebreaker, promoting the second-best option. This is observed in the Figure, where the answer can change over a

Model	**Dataset**	**Anger**		**Disgust**		**Fear**		Happiness		Sadness		**Surprise**	
		Acc	Layer	Acc	Layer	Acc	Layer	Acc	Layer	Acc	Layer	Acc	Layer
Llama-3.1-8B	crowd	0.91	30	0.945	31	0.92	27	0.96	29	0.965	25	0.855	22
	crowd_hidden	0.925	25	0.925	30	0.93	20	0.955	24	0.95	29	0.84	22
	dialogue	0.892	25	0.835	27	0.978	28	0.955	31	0.927	25	0.705	20
	twitter	0.865	30	-	-	0.725	30	0.755	28	0.745	25	0.68	29
	llm	0.995	21	0.995	20	1	16	1	19	1	22	0.95	15
	Pairwise Accuracy scores using the reading vectors from the LLM dataset												
	crowd	0.91	30	0.97	29	0.95	20	0.97	31	0.935	19	0.86	20
	crowd_hidden	0.925	25	0.935	29	0.965	25	0.95	22	0.94	22	0.865	21
	dialogue	0.897	27	0.883	20	0.956	20	0.93	21	0.921	19	0.77	20
	twitter	0.86	30	-	-	0.775	26	0.755	31	0.79	22	0.76	22
Gemma-2-9B	crowd	0.945	36	0.92	40	0.975	30	0.975	27	0.95	36	0.895	31
	crowd_hidden	0.93	35	0.95	34	0.945	29	0.965	25	0.93	30	0.88	32
	dialogue	0.918	34	0.864	30	1	30	0.955	34	0.969	34	0.88	29
	twitter	0.77	33	-	-	0.765	30	0.75	29	0.8	36	0.695	33
	llm	1	29	1	29	0.995	26	1	22	1	27	0.985	25
	Pairwise Accuracy scores using the reading vectors from the LLM dataset												
	crowd	0.96	35	0.93	34	0.97	33	0.975	29	0.96	27	0.92	32
	crowd_hidden	0.93	39	0.97	28	0.945	30	0.955	23	0.94	33	0.885	31
	dialogue	0.907	33	0.883	34	1	30	0.935	38	0.937	35	0.89	32
	twitter	0.825	39	-	-	0.765	30	0.745	27	0.765	35	0.745	33

Table 2: Maximum Pairwise Accuracy (Acc) and the associated layer for the reading vectors in detecting the emotion across datasets. Shows functionality and shared emotion reading representations. Prompt without explicit emotion (P4) used here.

range of JSD values.

• **Is the answer option determined in earlier layers?** We use Logit Lens (nostalgebraist, 2020) to take the hidden representations from the model and calculate the probability assigned to only the option tokens. Figure 8 shows the average probability values assigned to the correct option and the predicted option in the direct and CoT case for each layer. Interestingly, the Llama model begins considering option tokens in its later layers. In contrast, Gemma briefly considers the options in its early layers and then revisits them in the final layers. This indicates a difference in how the models process the input and what they focus on.

• **Effect of passing hint**: Results in Table 3 indicate that the emotion hint seems more effective in directing the model towards the right answer. However, it is crucial to note that when the hint is present, apart from the Llama-8B hint case, other models do not improve with CoT. In some cases, they score lower than with direct prompting. To test if the hint is actually incorporated into the CoT, we use the CoT generated as part of the prompt in place of the hint to generate the answer directly. The results shown in column Hint Influence Analysis - Acc. (CoT-H) columns show scores similar to the Direct prediction case for the Llama model, while Gemma showed lower scores. This indicates that the Llama model successfully incorporated the hint into its CoT trace while Gemma might have attended to the hint directly, without feeling the need to incorporate it into its reasoning trace. Hence, we rerun the prompt with additional instructions to explicitly incorporate the hint in its reasoning. The results shown in column Acc. (CoT-H2) show improvement in scores in Gemma when the hint is the emotion label, with only slight improvement when the hint is the cause. Overall, this experiment highlights the inherent limitation in emotion reasoning ability (with CoT) in these models, as they are unable to fully incorporate or connect the hint with the scenario.

6 Findings

From the experiments, we find that a) Emotion concepts are implicitly represented by the model within its representations (even when the instruction prompt does not mention it). The instruction style in the prompt affects emotion representations in early layers. Longer instructions with few-shot prompts show lower accuracy in early layers but get higher accuracy in later layers. b) There is high similarity in the emotion reading vectors obtained from different datasets. And these vectors are functionally equivalent in the later layers of the model, which we show through similar accuracy scores with the vectors from LLM dataset on other datasets and the vectors obtained from the dataset itself. LLMs are exposed to diverse data, including emotions, in their training phase, which might help them learn a combined average representation of emotions. This can be later used for emotion identification tasks. c) CoT style reasoning provides low gains in accuracy and shows a high overlap with the direct prompt predictions. However, the average entropy over the possible choices is lower with CoT compared to direct prompting, indicating that the model may use CoT to become more confident

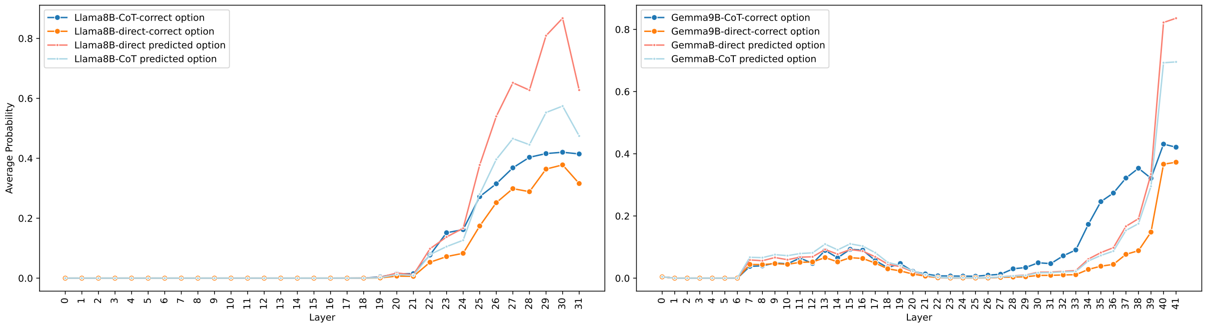


Figure 8: Layer wise Probability assigned to options

Model	Prompt Type	Index Prediction			Augmented Pred.		Hint Influence Analysis		Agreement of CoT and direct		Disagreement of CoT and direct	
		Acc. (Direct)	Acc. (CoT)	Missed (CoT)	Acc. (CoT-Aug)	Missed (Aug)	Acc. (CoT-H)	Acc. (CoT-H2)	Same Preds.	Same (Correct)	DW-CC	DC-CW
Llama-3.2-1B	direct/CoT	0.185	0.165	0.2	0.215	0.02	-	-	0.35	0.09	0.125	0.095
	hint_cause	0.16	0.21	0.16	0.27	0.005	0.225	0.195	0.27	0.09	0.18	0.07
	hint_cause_first	0.225	0.15	0.27	0.225	0.01	0.195	0.23	0.38	0.12	0.105	0.105
	hint_emo	0.27	0.235	0.16	0.3	0.01	0.255	0.255	0.31	0.15	0.15	0.12
	hint_emo_first	0.49	0.305	0.22	0.44	0.02	0.3	0.32	0.425	0.32	0.12	0.17
Llama-3.1-8B	direct/CoT	0.265	0.425	0.015	0.43	0.01	-	-	0.525	0.21	0.22	0.055
	hint_cause	0.33	0.62	0.005	0.62	0	0.605	0.58	0.48	0.28	0.34	0.05
	hint_cause_first	0.435	0.645	0.015	0.65	0	0.64	0.59	0.59	0.39	0.26	0.045
	hint_emo	0.815	0.855	0.005	0.855	0.005	0.81	0.765	0.815	0.76	0.095	0.055
	hint_emo_first	0.845	0.855	0.005	0.855	0	0.78	0.84	0.805	0.775	0.08	0.07
Gemma-2-2B	direct/CoT	0.25	0.22	0	0.22	0	-	-	0.63	0.16	0.06	0.09
	hint_cause	0.395	0.35	0	0.35	0	0.39	0.4	0.665	0.275	0.075	0.12
	hint_cause_first	0.41	0.385	0	0.385	0	0.365	0.39	0.635	0.3	0.085	0.11
	hint_emo	0.73	0.52	0	0.52	0	0.375	0.64	0.665	0.505	0.015	0.225
	hint_emo_first	0.705	0.495	0	0.495	0	0.39	0.575	0.675	0.475	0.02	0.23
Gemma-2-9B	direct/CoT	0.435	0.455	0	0.455	0	-	-	0.775	0.375	0.08	0.06
	hint_cause	0.585	0.59	0	0.59	0	0.545	0.59	0.795	0.525	0.065	0.06
	hint_cause_first	0.64	0.635	0.005	0.635	0.005	0.555	0.575	0.815	0.585	0.05	0.055
	hint_emo	0.82	0.74	0	0.74	0	0.585	0.7	0.885	0.73	0.01	0.09
	hint_emo_first	0.745	0.71	0	0.71	0	0.59	0.66	0.82	0.665	0.045	0.08

Table 3: Impact of CoT and Hints on Accuracy of emotion classification
Header abbreviations: **Acc. (Direct)**: Direct Prediction Accuracy. **Acc. (CoT)**: Chain-of-Thought Accuracy. **Missed**: Missed values for the corresponding accuracy column. **Acc. (CoT-Aug)**: Accuracy for CoT considering the selected option text, where index was not generated. **CoT-H**: Passing the CoT generated with the hint as part of the prompt and getting the prediction **CoT-H2**: The CoT was generated with a prompt explicitly asking to include the hint in its reasoning. **DW-CC**: Direct Wrong, CoT Correct. **DC-CW**: Direct Correct, CoT Wrong.

Model	Prompt	Accurate	Entropy scores			Top2-diff	JSD
			Mean	Median	Std		
Llama-3.1-8B	**Direct**	True	1.381	1.548	0.756	0.495	0.228
		False	1.56	1.663	0.608	0.43	
	CoT	True	0.184	0.033	0.297	0.942	
		False	0.258	0.043	0.427	0.867	
Gemma-2-9B	**Direct**	True	0.378	0.208	0.403	0.817	0.125
		False	0.486	0.294	0.436	0.7534	
	CoT	True	0.1	0.023	0.166	0.956	
		False	0.224	0.075	0.337	0.878	

Table 4: Aggregated Entropy scores over option indexes

in its predicted answer, regardless of its accuracy. Llama-8B showed gains with CoT. This can be attributed to its initial higher entropy, making it more susceptible to CoT reasoning. Gemma-9B, with relatively lower entropy in initial predictions, shows lower gains with CoT. (d) Providing the cause for the emotion in the form of hints helps the model generate more accurate CoT traces that are helpful to increase the accuracy of the model. This indicates that while CoT has the potential to be improved if directed in the right direction, it is not inherently capable of reaching that goal on its own.

7 Conclusion

In this paper, we present a unified understanding of how emotion concepts are represented in the model. We used supervised and unsupervised probing techniques to show the variation in the layers that show the presence of emotion concepts. This variation is an influence of the instruction prompt and the clarity with which emotion is expressed in the input data. We show the presence of intrinsic emotion reading vectors that are similar across datasets and can be used interchangeably, revealing their foundational nature. While LLMs have learned a representation of emotions, they are unable to leverage it for accurate reasoning when evaluated with CoT. We observe CoT mostly generates reasoning traces to increase it confidence in its original answer, especially when the model is confident in it. This motivates the need for methods that can leverage the implicit emotion representations to improve explicit reasoning capabilities of LLMs.

Limitations

We considered the popularly studied six-emotion classification by (Ekman et al., 1987) to ensure consistency across the dataset. While we demonstrate the presence of emotion concept spaces for these emotions, the representation directions for other, more fine-grained emotions may not be as clearly evident due to overlap and insufficient representation. We observe a difference in representation level accuracy between the 1B and 8B models, suggesting that larger models may have similar or better emotion representations. We experiment with models of up to 9B parameters due to computational budget constraints, which are crucial for white-box interpretability methods like probing.

References

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

P Ekman, W V Friesen, M O'Sullivan, A Chan, I Diacoyanni-Tarlatzis, K Heider, R Krause, W A LeCompte, T Pitcairn, and P E Ricci-Bitti. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *J. Pers. Soc. Psychol.*, 53(4):712–717.

Gemma Team. 2025. Gemma 2: A family of open models from google. `https://storage.googleapis.com/deepmind-media/gemma/gemma-2-report.pdf`. Technical report.

Yihuai Hong and Aldo Lipani. 2024. Interpretability-based tailored knowledge editing in transformers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3847–3858, Miami, Florida, USA. Association for Computational Linguistics.

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, Vienna, Austria. Association for Computational Linguistics.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Daria Lioubashevski, Tomer M. Schlank, Gabriel Stanovsky, and Ariel Goldstein. 2025. Looking beyond the top-1: Transformers determine top tokens in order. In *Forty-second International Conference on Machine Learning*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Aishwarya Maheswaran, Caslon Chua, and Maunendra Sankar Desarkar. 2025. Probing the inherent ability of large language models for generating empathetic responses. In *2025 IEEE Swiss Conference on Data Science (SDS)*, pages 32–39.

nostalgebraist. 2020. Interpreting GPT: The logit lens. *LessWrong*.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6124–6142, Vienna, Austria. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander F. Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations. *Cogn. Comput.*, 13(5):1317–1332.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. 2024. Does large language model contain task-specific neurons? In *Proceedings of the 2024 Conference on*

Empirical Methods in Natural Language Processing, pages 7101–7113, Miami, Florida, USA. Association for Computational Linguistics.

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, Vienna, Austria. Association for Computational Linguistics.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.

Hugo Touvron, A. Grattafiori, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint*.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Appendix: Dataset Details

All the datasets we used are publicly available to be used for research purposes. The LLM dataset (Zou et al., 2025) and EmoBench dataset (Sabour et al., 2024) have MIT license. The Twitter Emotions dataset was taken from (Palma et al., 2025) which has the Apache-2.0 license. All datasets are in English language.

- Twitter Emotions Dataset (Saravia et al., 2018) - Collected using the Twitter API and annotated automatically using the relevant hashtags. We use the same six emotion reduced data split as in (Palma et al., 2025).

- LLM-generated Dataset - Introduced in (Zou et al., 2025), this dataset is designed to be in the second person, with the intent to invoke a particular emotion in the reader. This dataset is vital because it does not explicitly contain any emotion keywords and tests the model's event-to-emotion association.

- RECCON Dataset (Poria et al., 2021) - We consider the Daily Dialog split of this dataset. This dataset is considered for comparison with emotion representation in dialogue settings. The emotion labels were skewed towards the 'joy' class, so a subset of this dataset was selected to balance the emotion class distribution for the classification task.

- Crowd-eVent Dataset (Troiano et al., 2023) - Consists of vignetts expressing 13 different emotion classes, with self-reported emotion annotations. We consider six emotion classes from here to make it comparable with the other datasets. This dataset has an additional version with explicit emotion words masked out from the text. We run experiments on both these versions. This dataset was also used in (Tak et al., 2025).

- EmoBench Dataset (Sabour et al., 2024) - The Emotion Understanding (EU) split of this dataset consists of Scenarios, typically involving more than one person, and asks the emotion expressed by a particular person through a list of emotion choices (MCQ - 6 option format). The scenarios presented here are categorized to depict complex emotions, including transitions between emotions, mixed emotions, unexpected outcomes, and cultural values. The emotion choices are fine-grained (38 possibilities with 3 additional choices for lack of emotion) and allow multiple emotions within a single option. It also has a cause associated with each scenario that explains why the subject might express the emotion. We selected this dataset to explore the limitations of emotion comprehension in LLMs and whether a Chain of Thought (CoT) style of prompting enhances the model's ability to comprehend emotions.

Dataset	Train	Test
LLM	1000 -> sadness: 183, surprise: 167, disgust: 164, happiness: 162, anger: 162, fear: 162	250 -> sadness: 45, anger: 41, disgust: 41, happiness: 41, surprise: 41, fear: 41
Crowd-eVent	2640 -> fear: 440, disgust: 440, joy: 440, surprise: 440, anger: 440, sadness: 440	660 -> fear: 110, disgust: 110, joy: 110, surprise: 110, anger: 110, sadness: 110
Twitter	6000 -> joy: 2010, sadness: 1750, anger: 810, fear: 726, love: 490, surprise: 214	2000 -> joy: 695, sadness: 581, anger: 275, fear: 224, love: 159, surprise: 66
Dialogue	951 -> anger: 200, joy: 200, surprise: 200, sadness: 200, disgust: 106, fear: 45	463 -> anger: 100, joy: 100, surprise: 100, sadness: 100, disgust: 34, fear: 29
EmoBench	200 scenarios with 6 diverse emotion options in MCQ format	
For the unsupervised probing experiments, we sample at most 200 examples per emotion. We use shuffled pairs from the training data as the test set, following the same approach as (Zou et al., 2025)		

Table 5: Dataset sizes

B Appendix: Prompt and Model Details

The exact prompts used for the experiments are listed here for reference. The terms in { } are replaced with appropriate values from the dataset. All our experiments are run on NVIDIA-6000 Ada GPU. The time to obtain the generations with batch size of 1 ranged from 1 to 3 hours, depending on the dataset size. All LLMs used were implemented using the Hugging Face framework. The CoT experiments performed on 8B and 9B parameter models took upto three hours The scores are reported over a single run of the model over the dataset with temperature as 0.1 to reduce variations. For the CoT experiments, we set the max new tokens to 425.

B.1 Supervised Probing Experiment

P0 - Direct data, with no instruction to generate the emotion. This was the prompt setup used in (Palma et al., 2025). We include this prompt to understand if emotion representations occur implicitly in models. P1 - Instruction asking to detect the emotion expressed in the input:

Prompt P1

What is the emotion expressed in the following context? Context: `{input_text}`

P2 - Similar to P1 with few shot examples similar to the setup used by (Tak et al., 2025).

Prompt P2

What are the inferred emotions in the following contexts? Context: My first child was born. Answer: joy Context: My dog died last week. Answer: sadness Context: `{input_text}` Answer:

P2-d - For the dialogue dataset, we modify the few-shot examples to have a dialogue context.

Prompt P2-d: for Dialog Data

What is the inferred emotion in the following context? Context: A: Today was a memorable day. B: Oh, what happened? A: My first child was born. Emotion of A: joy Context: B: I have been feeling low lately. A: Do you mind sharing what has been bothering you lately? B: My dog died last week. Emotion of B: sadness Context: `{dialogue}`

P1-long - Ablation prompt to account for influence of prompt length when comparing P1 and P2. Specifically P2 consists of 40 (41) tokens and P1 long consists of 41 (42) tokens for Llama and Gemma models respectively.

Prompt P1 long

Please infer the primary emotion that is being communicated or suggested within the following context. Answer based on the overall tone and expressed feelings present in the text as a whole. Context: `{input_text}`

B.2 Unsupervised Probing Experiments

For the unsupervised probing experiment, we use the following prompts to obtain the hidden states to compare the effect of the instruction prompt on the emotion representation.

> **Prompt P3**
>
> Consider the {emotion} of the following scenario: Scenario: {scenario} Answer:

This is the same prompt used in (Zou et al., 2025), which we refer to as the prompt 'with explicit emotion'.

> **Prompt P4**
>
> What is the inferred emotion in the following scenario: Scenario: {scenario} Answer:

This is the prompt we use to obtain the emotion reading vectors and the results reported in Table 2. We refer to this prompt as 'without explicit emotion'.

> **Prompt P5**
>
> What is the inferred emotion in the following context? Speaker {speaker}: {scenario} Emotion of {speaker}: Answer:

Here the speaker is randomly assigned A or B to mimic a dialogue setting. This prompt along with P3 and P4 was used to compare the effect of prompt instruction in the emotion representation (shown in Figure 3).

B.3 CoT Experiments

For the CoT experiments with the EmoBench dataset, we primarily use these two prompts:

> **Prompt P6: Direct Prompt**
>
> Identify the emotion experienced by the given subject in the scenario. And then choose the closest emotion from the choices. Output only the appropriate choice number nothing else. Scenario: {scenario}, Subject: {subject} Choices: {indexed_choices}. Output:

> **Prompt P7: CoT Prompt**
>
> Identify the emotion experienced by the given subject in scenario. And then choose the closest emotion from choices. Think step by step to arrive at the answer, as if the choices did not exist. Do not mention, evaluate or eliminate the provided choices in your generation. Scenario: {scenario}, Subject: {subject} Choices: {indexed_choices}. Always end with 'selected choice index =' followed by the chosen choice index only.

We also experiment with providing hints along with the instructions. The hint location was placed before the scenario is provided (earlier) and after the choices ('late').

> **Prompt P8: Hint-early shown in the CoT prompt**
>
> Identify the emotion experienced by the given subject in scenario. And then choose the closest emotion from choices. Think step by step to arrive at the answer, as if the choices did not exist. Do not mention, evaluate or eliminate the provided choices in your generation. Hint: {hint}. Scenario: {scenario}, Subject: {subject} Choices: {indexed_choices}. Always end with 'selected choice index =' followed by the chosen choice index only.

> **Prompt P9: Hint-late shown in Direct prompt**
>
> Identify the emotion experienced by the given subject in scenario. And then choose the closest emotion from choices. Output only the appropriate choice number nothing else. Scenario: {scenario}, Subject: {subject} Choices: {indexed_choices}. Hint: {hint}. Output:

C Appendix: Importance of Last Token

LLMs have layer and component-level representations for each token in the input. In general, the last token's representation is considered for the interpretability analysis as it is found to contain information relevant for the task and directly influences the next token to be predicted (Hong and Lipani, 2024) (Lioubashevski et al., 2025). (Tak et al., 2025) predicted emotions using the k^{th} last token (with $k = 1, 2, ..., 5$) as input. It was observed that the performance gradually increases as

k goes from 5 to 1 with the best results with the last token, emphasizing its usefulness. Pooling-based methods like concatenating the *min*, *mean*, and *max* of token representations were also found to perform well in (Palma et al., 2025).

D Appendix: Unsupervised Probing Method Details

The data is preprocessed into pairs of the form: $D^e_{train} = (x^+_i, x^-_i)^N_{i=1}$ where x^+_i is a sample expressing the emotion of interest e and x^-_i is a sample expressing any other emotion. Each sample in the data pairs, along with the instruction prompt, is passed to the model to obtain the last token hidden representations, denoted as $h^+_i = M(x^+_i)[-1, :]$. The difference of the representations for each pair: $\delta_i = h^+_i - h^-_i$, gives us a set of N difference vectors $\Delta^e = \delta_1, \delta_2, ...\delta_N$. A PCA model is fit on this data (Δ^e). The resulting first principal component is considered the emotion reading vector $d^{e'} = PCA_1(\Delta^e)$. We then calculate the scalar projection $p^{+/-}$ of the individual hidden representations $h^{+/-}$ with $d^{e'}$ (denoted as $S(h, d) = (h.d)/||d||_{L2}$). The sign ($\sigma^e$) associated with $d^{e'}$ is obtained by averaging the difference in the projection values across the training set. This represents the frequency of cases where the scalar projection of the sample with the target concept is higher or lower than the random sample in its pair, giving us + sign when its greater on average and – sign when its lower on average. Mathematically its represented as: $\sigma^e = sign(\sum^N_{i=1}[I(p^+_i = max(p_i)) - I(p^+_i = min(p_i))])$. Including the sign with the previously calculated emotion reading gives us the final emotion reading vector $d^e = \sigma^e.d^{e'}$. We calculate the pairwise accuracy score by first computing the scalar projections of the samples in the test set. For a data pair (x^+_j, x^-j) in the test set giving last token hidden representations as (h^+_j, h^-_j) the scalar projection is determined as $p^+_j = S(h^+_j, d^e)$ and $p^-_j = S(h^-_j, d^e)$. The prediction is considered correct if $p^+_j > p^-_j$, $Accuracy^e = \frac{1}{K}\sum I(p^+_j > p^-_j)$

E Appendix: Additional Results

E.1 Temperature 0 results

We originally used a temperature of 0.1 as a near-deterministic setting for emotion-option selection, avoiding undesirable behaviors such as repetition loops or rigid max-token bias at temperature 0. Results using temperature = 0 were nearly identical in most prompting experiments (typically <2% difference). Overall in 40 experimental setups, only 4 experiments saw a change >5% (refer Table 6). This was mostly observed for the Llama-3.2-1B-Instruct model, in the experimental setting involving hints, particularly when the hint explicitly contained the target emotion.

E.2 Probing Results for other Models

Please refer Figures 9, 10 and 11 for Supervised Probing results for models Llama-3.2-3B, Gemma-2-2B, Gemma-2-9B.

E.3 Activation Patching Results

Please refer Figures 12 and 13 for Activation Patching results for models Llama-3.1-8B and Gemma-2-9B.

E.4 Emotion Neuron Identification

To identify emotion specific neurons we utilize the entropy based approach introduced in (Tang et al., 2024) and used in (Lee et al., 2025). Overall we observe that around 100-300 neurons were identified at each layer for different emotions across datasets (Refer Figures 14, 15,18). Taking the intersection of the identified neurons (common neurons) we observe that the frequency reduces to around 10-20 neurons at each layer (refer 16, 17). With the absence of the instruction prompt, the common emotion neurons further drops to less than 10 (refer Figure 19).

Figures 14 and 15 show the count of emotion-specific neurons identified for each dataset.

Figures 16 and 17 show the count of common emotion-specific neurons identified across dataset.

Figures 18 and 19 show the count of emotion-specific neurons identified for each dataset and the common neurons across datasets for Llama-3.1-8B Instruct model with prompt P0 (no instruction).

F Appendix: CoT Experiment Details and Examples

The Jensen Shannon Divergence metric (JSD) (Lin, 1991) is calculated as follows: $JSD(P, Q) = \frac{1}{2}(D_{KL}(P_{direct}, M) + D_{KL}(P_{CoT}, M))$ where $M = \frac{1}{2}(P_{direct} + P_{CoT})$. D_{KL} is the Kullback-Leibler Divergence calculated as $D_{KL} = \sum^5_{i=0} p_i.log_2(\frac{p_i}{M_i})$.

Please refer Table 7 for the examples.

Model	Prompt	Temp = 0			Temp = 0.1		
		Acc (Dir)	Acc (CoT)	Missed	Acc (Dir)	Acc (CoT)	Missed
Llama-3.2-1B-Instruct	direct/CoT	0.175	0.170	0.190	0.185	0.165	0.200
	hint_cause	0.180	0.180	0.125	0.160	0.210	0.160
	hint_cause_first	0.255	0.225	0.230	0.225	0.150	0.270
	hint_emo	0.365	0.285	0.135	0.270	0.235	0.160
	hint_emo_first	0.460	0.320	0.195	0.490	0.305	0.220
Llama-3.1-8B-Instruct	direct/CoT	0.275	0.440	0.015	0.265	0.425	0.015
	hint_cause	0.340	0.590	0.015	0.330	0.620	0.005
	hint_cause_first	0.445	0.650	0.010	0.435	0.645	0.015
	hint_emo	0.820	0.845	0.000	0.815	0.855	0.005
	hint_emo_first	0.835	0.835	0.000	0.845	0.855	0.005
gemma-2-2b-it	direct/CoT	0.260	0.265	0.000	0.250	0.220	0.000
	hint_cause	0.395	0.360	0.000	0.395	0.350	0.000
	hint_cause_first	0.415	0.355	0.000	0.410	0.385	0.000
	hint_emo	0.725	0.520	0.000	0.730	0.520	0.000
	hint_emo_first	0.710	0.470	0.000	0.705	0.495	0.000
gemma-2-9b-it	direct/CoT	0.440	0.435	0.000	0.435	0.455	0.000
	hint_cause	0.585	0.600	0.005	0.585	0.590	0.000
	hint_cause_first	0.645	0.635	0.000	0.640	0.635	0.005
	hint_emo	0.820	0.720	0.000	0.820	0.740	0.000
	hint_emo_first	0.740	0.705	0.000	0.745	0.710	0.000

Table 6: Comparison of Accuracy and Missed Values at Temp 0 vs 0.1

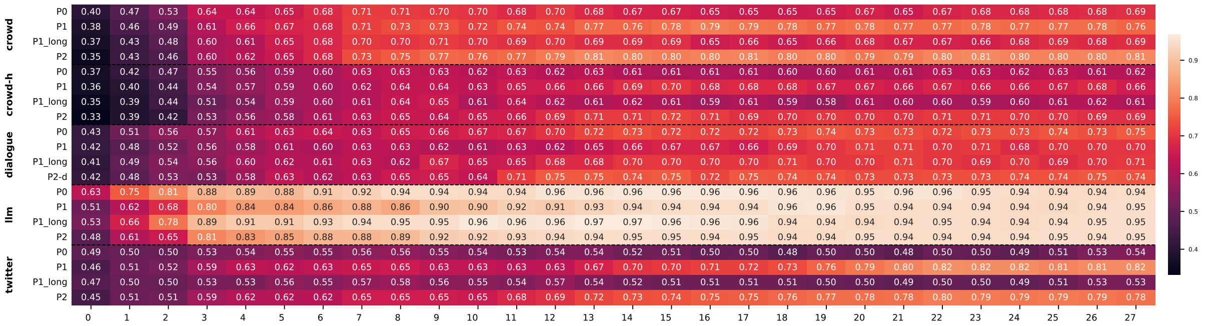


Figure 9: Probing Llama-3.2-3B 6-emotion results

G Appendix: AI use

We used AI in the form of grammatical error-correction tools to refine the text. We also used Gemini model for generating some basic code for the plots.

H Appendix: Probing Results for other MLP, Attention representations

We present the supervised probing results using representations from the MLP and the attention layers for completeness.

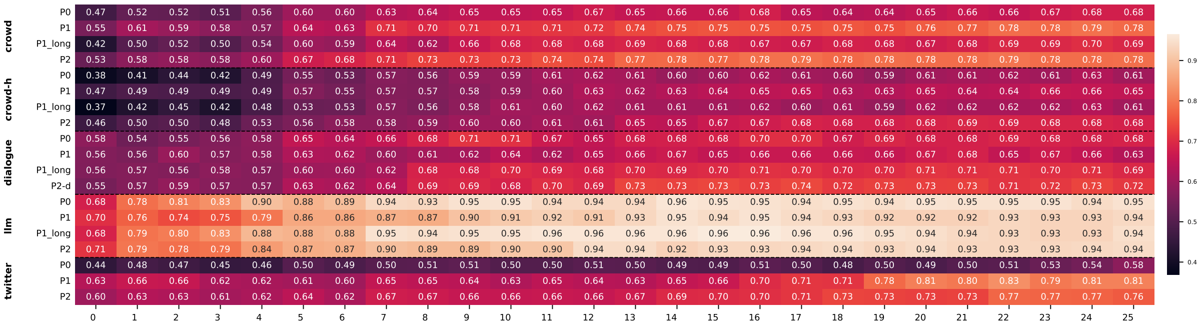


Figure 10: Probing Gemma-2-2B 6-emotion results

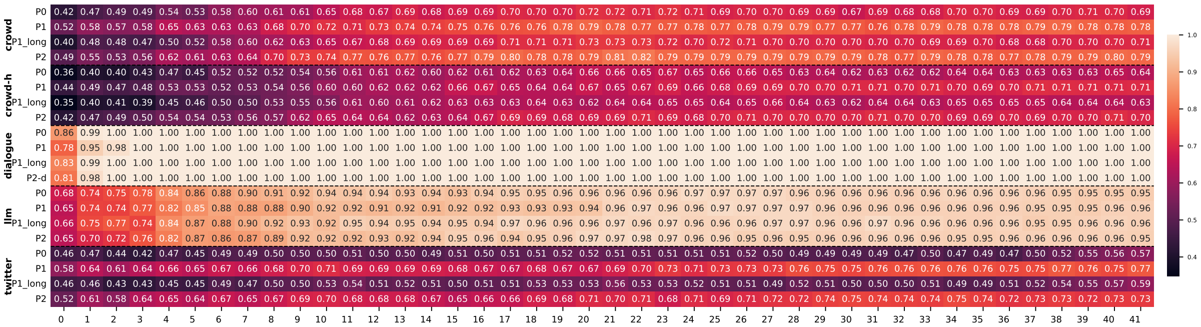


Figure 11: Probing Gemma-2-9B 6-emotion results

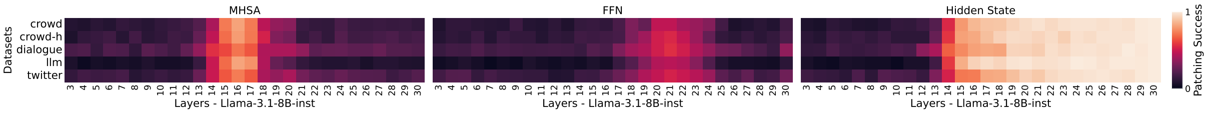


Figure 12: Patching Llama-3.1-8B 6-emotion results

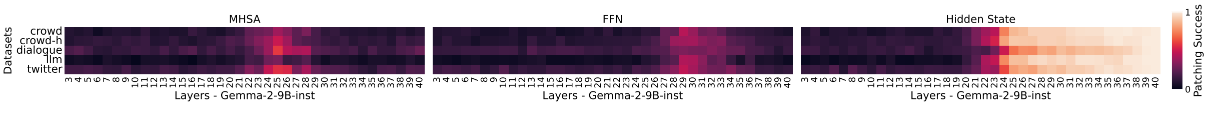


Figure 13: Patching Gemma-2-9B 6-emotion results

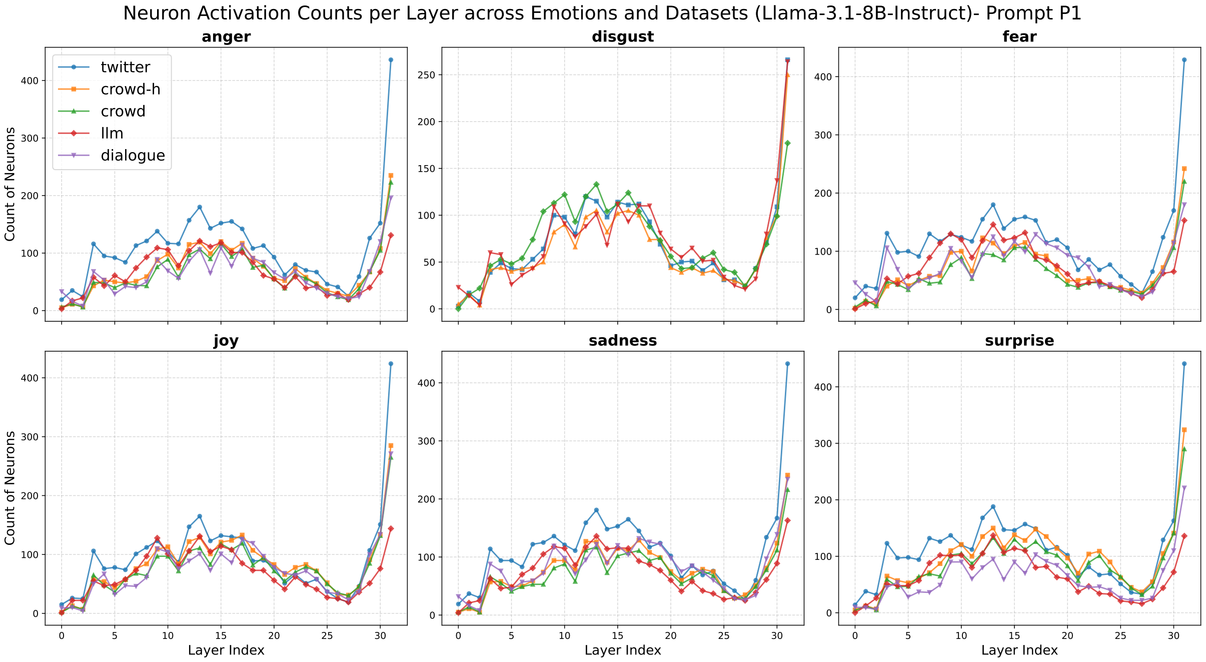


Figure 14: Dataset specific identified emotion neurons Llama-3.1-8B (Prompt P1)

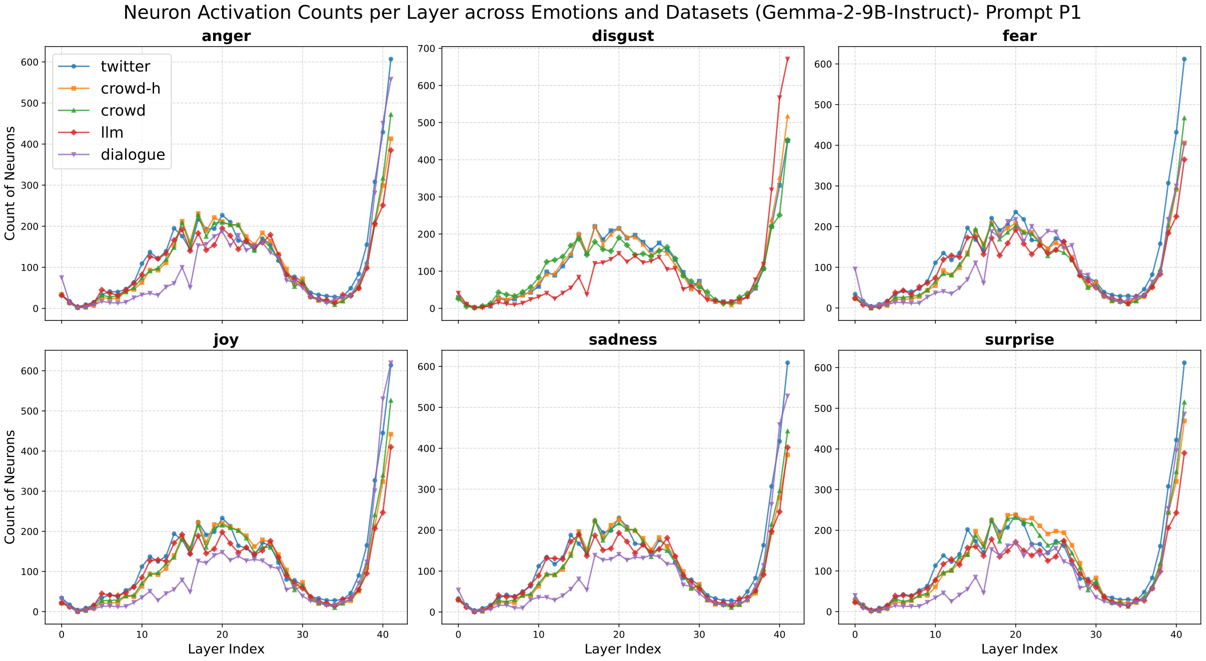


Figure 15: Dataset specific identified emotion neurons Gemma-2-9B (Prompt P1)

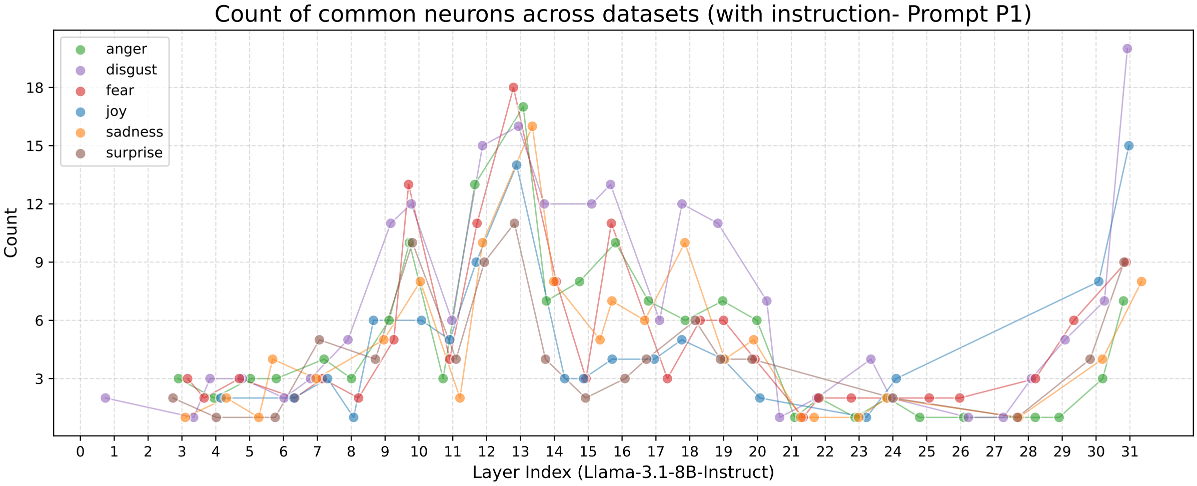


Figure 16: Frequency of emotion neurons common across datasets, Llama-3.1-8B (Prompt P1)

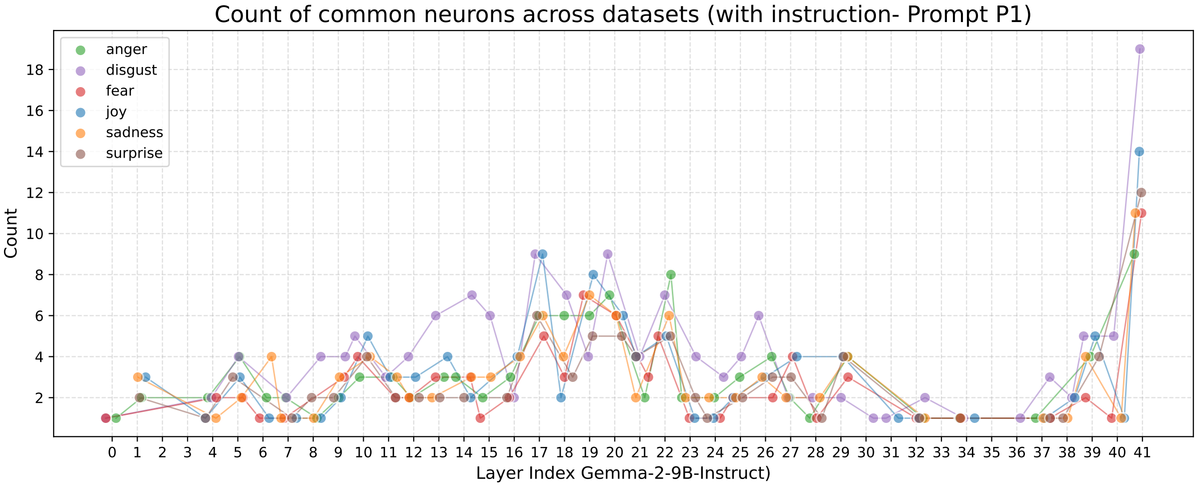


Figure 17: Frequency of emotion neurons common across datasets, Gemma-2-9B (Prompt P1)

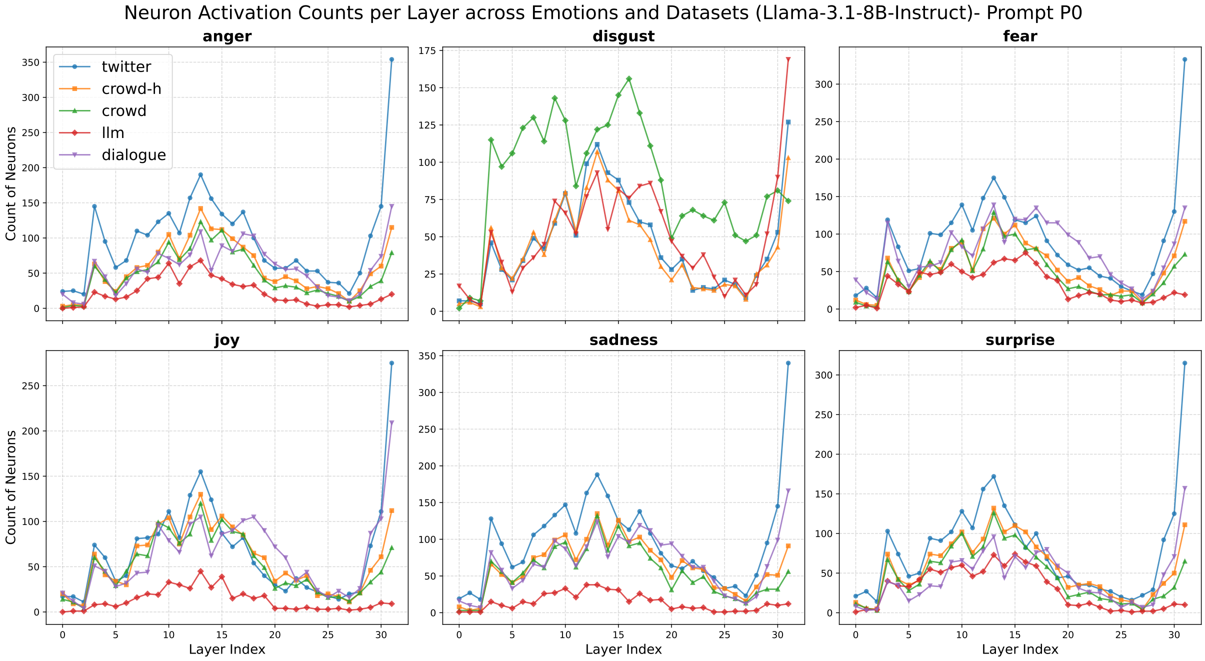


Figure 18: Dataset specific identified emotion neurons Llama-3.1-8B (Prompt P0)

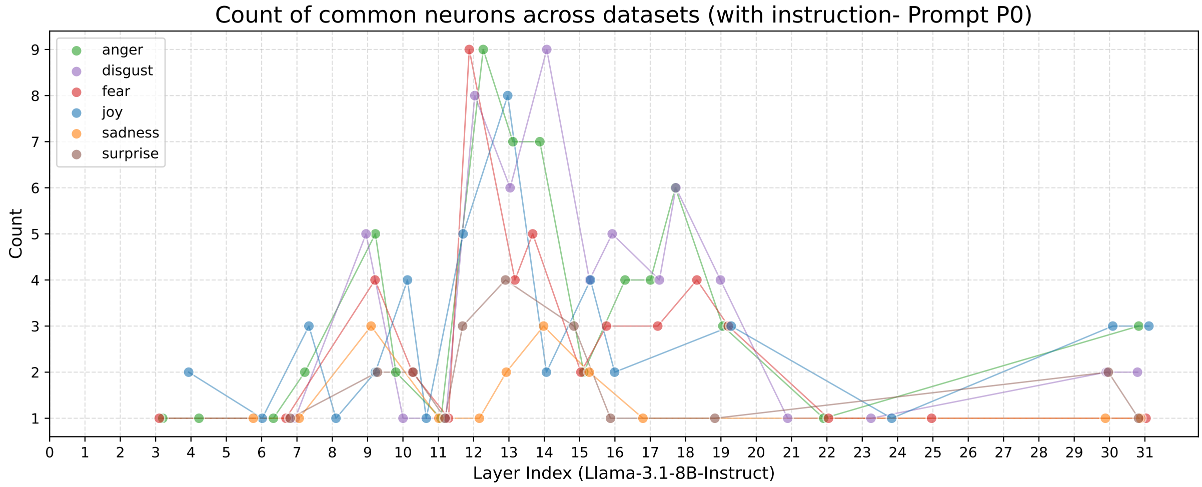


Figure 19: Frequency of emotion neurons common across datasets, Llama-3.1-8B (Prompt P0)

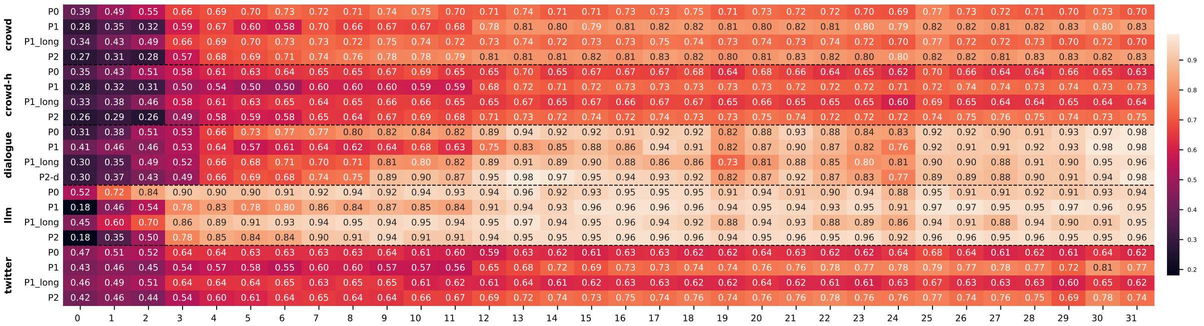


Figure 20: Probing Llama-3.1-8B attention layer

Scenario	**Options**	**Answer / Pred**
Chris absolutely loves drinking alcohol, regardless of its taste, and considers it his favorite hobby. His favorite type of alcohol is Whiskey. Yesterday, he went to his friend's bar to try out their new Whiskey-based cocktail. As Chris finished the drink in one big gulp, his nose wrinkled, and his eyes diminished. He then thanked his friend and asked for another cup. Subject: Chris	["Anger and Nervousness", "Joy and Nervousness", "Disapproval and Nervousness", "Joy and Disapproval", "Joy and Excitement", "Anger and Disgust"]	**Ground Truth:** 4: Joy and Excitement
Model: Llama-3.1-8B - CoT text To identify the emotion experienced by Chris, let's analyze the scenario step by step ... implies that he was excited to try something new and was likely looking forward to enjoying a drink that he thought he would like. 4. As Chris finished the drink in one big gulp, his nose wrinkled, and his eyes diminished. This reaction suggests that he was not pleased with the taste of the drink ... despite not liking the taste of the drink, he was still willing to continue trying it. This behavior suggests that he is not easily deterred by a bad taste and is willing to give the drink another chance... Chris is still enthusiastic about trying new drinks and is willing to continue trying the drink, which suggests a sense of joy. However, he is not pleased with the taste of the drink, which suggests a sense of disapproval.**selected choice index = 3**		**CoT:** 3 (Joy and Disapproval) **Direct:** 4
Model: Gemma-2-9B - CoT text Chris enjoys drinking alcohol and specifically whiskey. He tried a new whiskey cocktail and although his facial expressions suggest he didn't like the taste, he still asked for another. This indicates he enjoys the experience of drinking despite the taste not being to his liking. **selected choice index = 4**		**CoT:** 4 **Direct:** 4

Scenario	**Options**	**Answer / Pred**
During her visit to India, Sophia became close to Laila, her Indian Coworker, and was invited to attend her birthday party. As she wanted to stand out, she wore her favorite white dress. When Laila saw Sophia, she told her what her dress indicated. Subject: Sophia	["Embarrassment", "Joy", "Excitement", "Oblivious", "Gratitude", "Nervousness"]	**Ground Truth:** 0: Embarrassment
Model: Llama-3.1-8B - CoT text To identify the emotion experienced by Sophia, let's analyze the scenario step by step:1. Sophia became close to Laila, her Indian coworker, and was invited to attend her birthday party. This indicates that Sophia has a positive relationship with Laila and is likely feeling happy or excited about the invitation.2. Sophia wore her favorite white dress to stand out at the party. This suggests that Sophia is trying to make a good impression or feel confident.3. When Laila saw Sophia, she told her what her dress indicated. This implies that Laila might have shared some cultural or social significance of the dress, which could be embarrassing or awkward for Sophia if she wasn't aware of it.these points, it's likely that Sophia felt embarrassed when Laila shared the meaning of her dress, especially if she wasn't aware of it beforehand. This would be a natural reaction if she felt self-conscious or exposed.**selected choice index = 0**		**CoT:** 0 **Direct:** 4 (Gratitude)
Model: Gemma-2-9B - CoT text Sophia wanted to stand out at Laila's birthday party. She chose to wear her favorite white dress. Laila then told Sophia what her dress indicated. This suggests that Sophia's dress had a special meaning in Indian culture, perhaps related to a particular occasion or belief. Sophia likely felt a mix of emotions upon realizing this, possibly feeling a bit embarrassed or self-conscious that she hadn't been aware of the cultural significance. **selected choice index = 0**		**CoT:** 0 **Direct:** 0

Table 7: Example of EmoBench scenarios and CoT generations

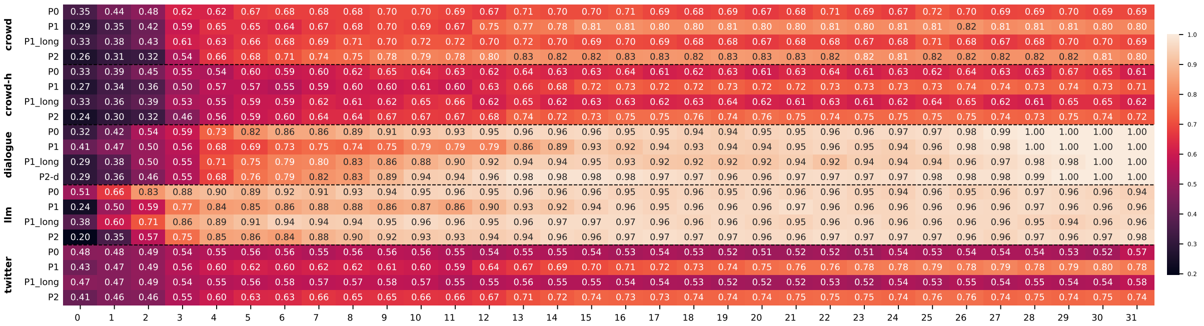


Figure 21: Probing Llama-3.1-8B MLP layer

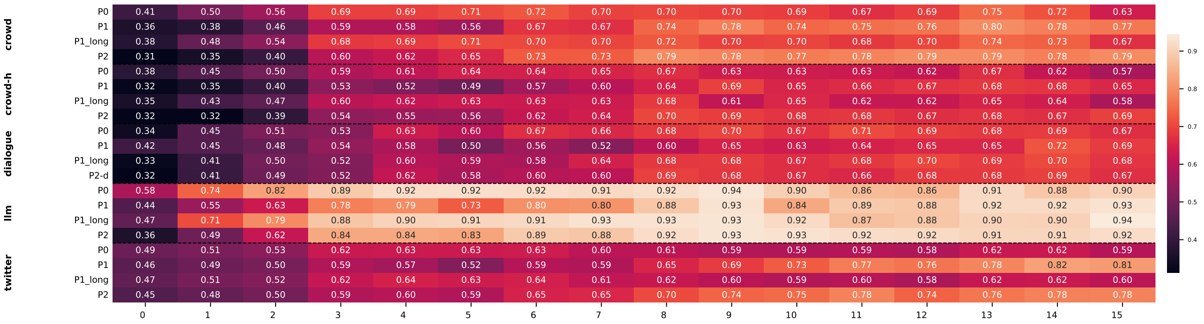


Figure 22: Probing Llama-3.2-1B attention layer

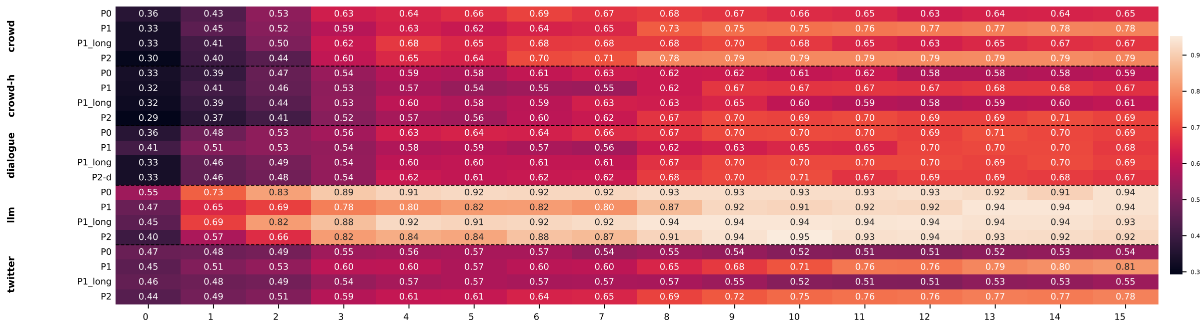


Figure 23: Probing Llama-3.2-1B MLP layer

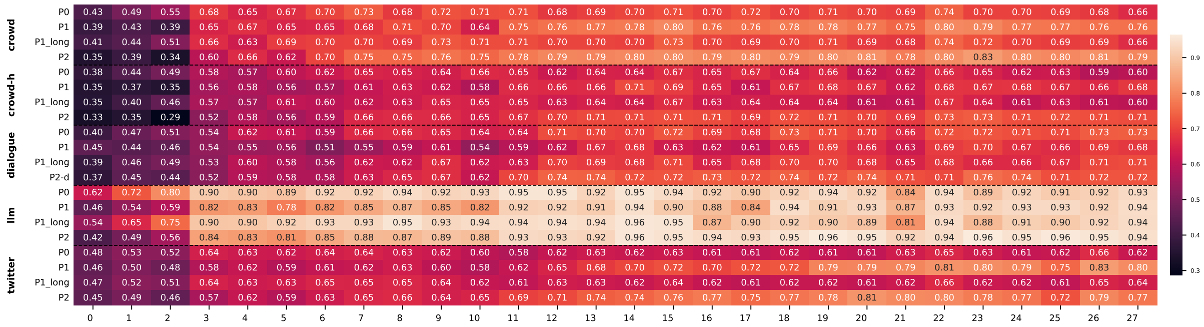


Figure 24: Probing Llama-3.2-3B attention layer

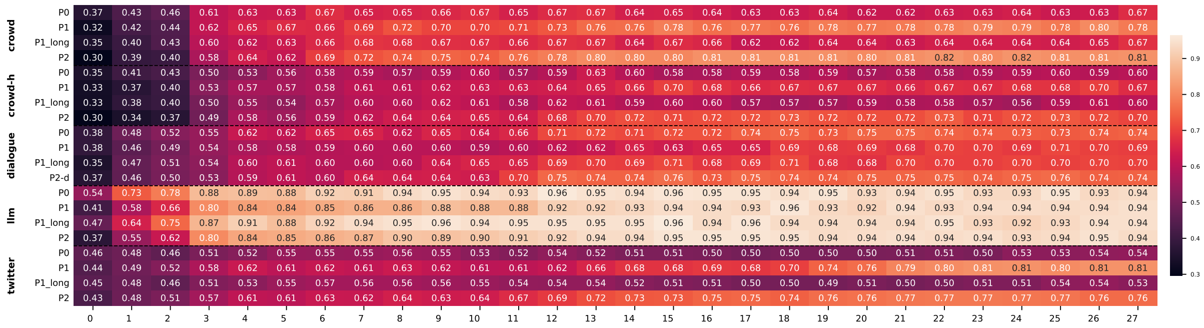


Figure 25: Probing Llama-3.2-3B MLP layer

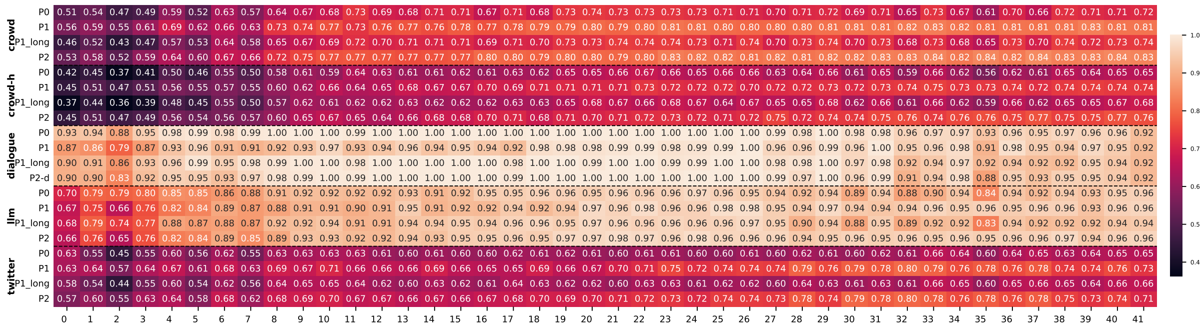


Figure 26: Probing Gemma-2-9B attention layer

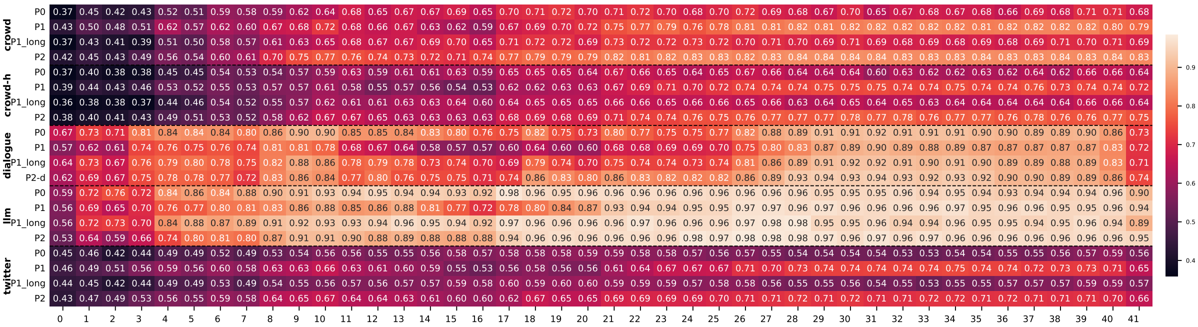


Figure 27: Probing Gemma-2-9B MLP layer

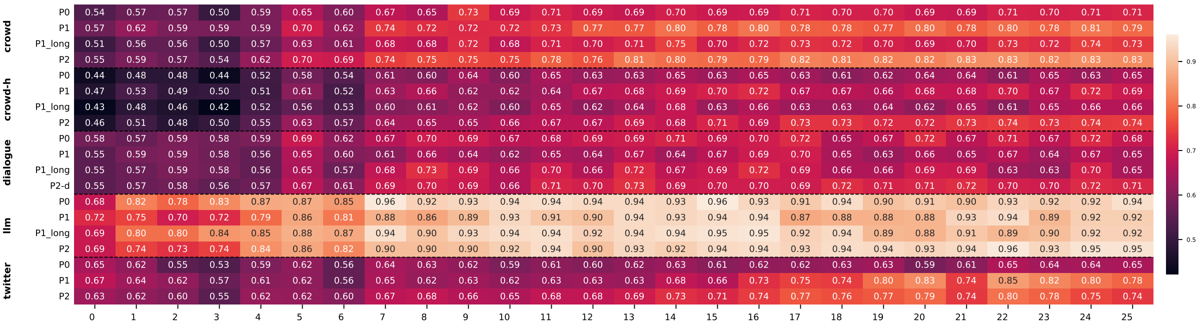


Figure 28: Probing Gemma-2-2B attention layer

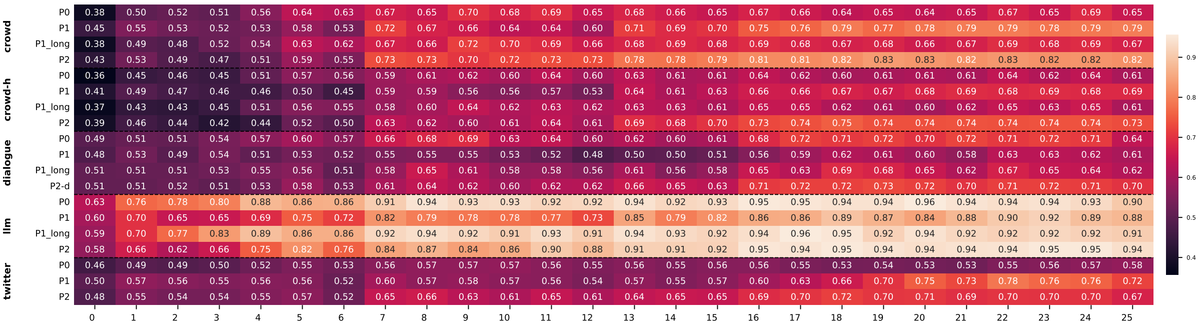


Figure 29: Probing Gemma-2-2B MLP layer