Mechanistic Interpretability of Emotion Inference in Large Language Models

Ala N. Tak[1,3,*], Amin Banayeeanzade[2,3,*], Anahita Bolourani[4], Mina Kian[3], Robin Jia[3], Jonathan Gratch[1,3],
[1]Institute for Creative Technologies, University of Southern California (USC),
[2] Information Sciences Institute, University of Southern California (USC),
[3]Department of Computer Science, University of Southern California (USC),
[4]Department of Statistics and Data Science, University of California, Los Angeles (UCLA)
Correspondence: antak@ict.usc.edu; banayeea@usc.edu, [*]Equal contribution

Abstract

Large language models (LLMs) show promising capabilities in predicting human emotions from text. However, the mechanisms through which these models process emotional stimuli remain largely unexplored. Our study addresses this gap by investigating how autoregressive LLMs infer emotions, showing that emotion representations are functionally localized to specific regions in the model. Our evaluation includes diverse model families and sizes, and is supported by robustness checks. We then show that the identified representations are psychologically plausible by drawing on cognitive appraisal theory—a well-established psychological framework positing that emotions emerge from evaluations (appraisals) of environmental stimuli. By causally intervening on construed appraisal concepts, we steer the generation and show that the outputs align with theoretical and intuitive expectations. This work highlights a novel way to causally intervene and control emotion inference, potentially benefiting safety and alignment in sensitive affective domains. *Code at:* `GitHub repo.`

1 Introduction

Large Language Models (LLMs) demonstrate remarkable capabilities in emotion recognition and reasoning tasks, occasionally surpassing human performance (Elyoseph et al., 2023; Tak and Gratch, 2024). Prior research primarily engages with LLMs as black boxes, utilizing zero-shot inference or in-context learning to gauge their performance on tasks such as emotion classification (Yongsatianchot et al., 2023; Broekens et al., 2023), emotional decision-making and situational appraisal (Tak and Gratch, 2023), emotional intelligence (Wang et al., 2023b), emotional dialogue understanding (Zhao et al., 2023), and generation of emotional text (Gagne and Dayan, 2023). However, there remains a limited understanding of *how* LLMs internally represent and process emotional information. Given LLMs' increasingly significant societal impact—spanning domains such as mental health (Sharma et al., 2023) and legal decision-making (Lai et al., 2024)—investigating these internal mechanisms is crucial.

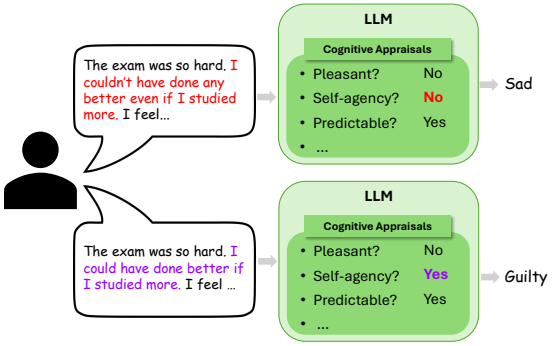

Figure 1: Emotion inference through latent appraisal-like mechanisms in LLMs. Given the description of a situation, the model leverages internal *appraisal* structures to recognize the emotion inferred from the context. For example, different perceptions of *self-agency* can distinguish between *guilt* and *sadness*.

Cognitive neuroscience uses functional localization approaches to identify specific brain regions responsible for particular functions and manipulate them by up/down-regulating neural activations in those regions. Akin to the shift from behaviorism to cognitive neuroscience in psychology—i.e. from treating the mind as a black box to studying brain-based cognitive processes—Mechanistic Interpretability (MI) allows for moving from black-box techniques (Casper et al., 2024), to a focus on the internal mechanics of LLMs (Bereska and Gavves, 2024). MI can offer a fundamental understanding of information processing and representation in LLMs, yielding insights into their inner-workings and offering new ways to control their reasoning (Li et al., 2021; Rai et al., 2024; Feng et al., 2025). Building on this line of research and by drawing inspiration from emotion theory in psychology, our work elucidates the inner workings of emotion processing in LLMs.

Findings of the Association for Computational Linguistics: ACL 2025, pages 13090–13120

We start by training linear classifiers on top of hidden representations to probe for regions where the strongest emotion-related activations occur. We provide evidence for functional localization of emotion processing and show that emotion-relevant operations are concentrated in specific layers, a consistent behavior across various model families and scales. We complement these findings by applying causal interventions, namely patching activations in the computation graphs, to identify essential components in neural representations (Conmy et al., 2023). As a result, we show that Multi-Head Self-Attention (MHSA) units in the mid-layers are predominantly responsible for shaping the LLM decision. To further corroborate this, we visualize attention patterns, revealing that MHSA units consistently attend to emotionally loaded tokens. Our findings are robust and not influenced by variations in prompt wording or formatting.

Additionally, we use the *appraisal theory* from psychology to shed light on the structure of LLMs' internal representations. According to appraisal theory (Frijda et al., 1989; Scherer et al., 1984; Smith and Ellsworth, 1985), people reason about emotional situations by forming *appraisal* judgments (see Figure 1 as an example). We analyze the structure of emotion representations in LLMs by conducting inference-time probing on appraisal concepts to show that representations are psychologically plausible. Moreover, by modulating latent appraisal concepts to promote/demote a particular appraisal dimension (e.g., promoting *self-agency*), we show that the resulting changes in the output emotion align with theoretical expectations (from *guilt* to *sadness*) (Marcinkevičs et al., 2024; Wu et al., 2024; Li et al., 2023).

Overall, our work extends existing MI methodologies by applying them to more ecologically valid, unstructured examples, moving beyond the common practice of analyzing simplified sentence structures (Wang et al., 2023a; Hanna et al., 2023). Furthermore, this work serves as an early step to bridge MI techniques with applications in psychological and cognitive domains, offering insights into the inner workings of LLMs in complex, socially relevant contexts.

2 Related Work

Appraisal Theory. Appraisal theory is a model that explains how peoples' emotions are a result of their evaluations of a situation (Lazarus, 1991). It provides a comprehensive framework for understanding the precursors of emotions (Smith and Kirby, 2011). Neuroscience studies build on this framework by manipulating cognitive appraisals and examining associated brain activity, linking specific brain regions to appraisal processes (Leitão et al., 2020; Kragel et al., 2024; Brosch and Sander, 2013). These methods can be extended to evaluate how LLMs understand emotions, and identify the mechanisms responsible for those evaluations.

Cognitive-Neuroscientific Alignment. There is a growing body of evidence demonstrating that neural network activations can reflect cognitive constructs traditionally studied in psychology and neuroscience. For example, transformer self-attention mechanisms are shown to correlate with human eye-tracking data during reading, suggesting that LLMs may learn attention patterns that reflect human cognitive attention (Bensemann et al., 2022; Eberle et al., 2022). LLMs also exhibit strong representational alignment with neural activity in the human language systems, indicating convergence between learned representations and brain-like processing (Aw et al., 2023), and word embeddings show alignment with fMRI activity during word association tasks (Kwon et al., 2024). Recent work even uses LLM-derived representations to predict neural activation patterns tied to high-level semantic categories like faces or places, offering scalable proxies for annotation in cognitive experiments (Liu et al., 2025). Other studies localize LLM units that align with cortical language areas (AlKhamissi et al., 2025) and reveal functional specialization patterns predictive of regional brain activity (Kumar et al., 2024).

These findings support the plausibility of mapping psychological or cognitive theories—such as appraisal theory—onto the internal structure of LLMs. Our work builds on this foundation by examining how appraisal mechanisms are internally encoded, bridging prior behavioral studies of appraisal-emotion mappings in LLMs (Tak and Gratch, 2024, 2023; Broekens et al., 2023; Yongsatianchot et al., 2023) with MI.

Mechanistic Interpretability. Probing is an MI technique that uses a simple model, called a "probe", to assess the internal representations across various layers in a model. As explained by Belinkov (2018), the groundwork for what we now refer to as probing relates back to earlier work evaluating trained classifiers on static word embeddings to predict linguistic features (Köhn, 2015;

Gupta et al., 2015), and classified hidden states of neural models (Ettinger et al., 2016; Kádár et al., 2017; Shi et al., 2016; Adi et al., 2017; Hupkes and Zuidema, 2018; Belinkov, 2022; Giulianelli et al., 2018). Probing is used across a variety of tasks (Hewitt and Liang, 2019; Tenney et al., 2019a,b; Peters et al., 2018; Clark et al., 2019; Belinkov, 2018; Conneau et al., 2018).

Activation patching (Heimersheim and Nanda, 2024), is a causal intervention used to identify if certain activations are important to the downstream task (Vig et al., 2020). By using patching, Meng et al. (2022) are able to localize where models store factual information. Patchscope, a method that extends on activation patching, is used to translate LLM representations into natural language (Ghandeharioun et al., 2024).

Yet another MI technique is generation steering. This method entails manipulating a model's activations to control the outputs (Rai et al., 2024; Todd et al., 2024). Geva et al. (2022) highlight the role of Feed-Forward Network (FFN) units in promoting concepts. To steer generation, they apply sub-updates promoting safety and are able to reduce the model's toxicity. Templeton et al. (2024) find that "clamping" on features can be used to control the models' output, steering the model's stated goals and biases for both desirable and undesirable outputs. Nanda et al. (2023) demonstrate that sequence models can have linear internal representations and that these representations can be used to manipulate the model's behavior. This method closely resembles those of Turner et al. (2023) and Lieberum et al. (2023).

3 Experimental Setup

Dataset and Prompt Design. We employ the crowd-enVENT dataset developed by Troiano et al. (2023), which comprises 6,800 emotional vignettes annotated with self-reported emotions among a list of 13 options and 23 self-rated appraisal variables, reflecting nuanced stimuli evaluations, including: *pleasantness/unpleasantness*, *self-agency/other-agency*, *predictability/suddenness*. Appendix A.1 presents more details on the dataset, including a detailed list of appraisal variables, along with the scales used for measurement.

To evaluate the model's ability to infer emotions from textual contexts, we design prompts that guide the model to predict the appropriate emotion as the next immediate output token, framing the task as a causal language modeling problem. Subsequently, we consider a classification problem and evaluate the model by inspecting the logits confined to the set of targeted emotion labels. The primary prompt template used in this study is shown in Figure 4.

Emotion attribution is inherently subjective, making it challenging to define a single ground truth label for each input, particularly given our fine-grained list of emotions. Thus, we focus on the correctly classified examples when inspecting each language model. In other words, we only analyze the samples for which the LLM and the human annotator agreed on the same label, totaling at least 2,700 samples among different language models (see Appendix A.3 for more details). This ensures that we assess tasks where the model performs reliably to understand the underlying mechanisms.

Model Architecture. To account for the impact of model scale and architectural variations, we evaluate a diverse set of model families and sizes, including Llama 3.2 1B Instruct and Llama 3.1 8B Instruct (Grattafiori et al., 2024), Gemma 2 2B Instruct and Gemma 2 9B Instruct (Team et al., 2024), OLMo 2 7B Instruct and OLMo 2 13B Instruct (OLMo et al., 2024), Phi 3.5 mini Instruct and Phi 3 medium-Instruct (Abdin et al., 2024), and Ministral 8B Instruct and Mistral 12B Nemo Instruct (MistralAI, 2024) (see Appendix A.2 for more details). Some detailed analyses, robustness tests, and appraisal concept interventions are exclusively conducted on Llama 3.2 1B to manage computational resources effectively.

4 Notations and Preliminaries

Prior research suggests that both MHSA and FFN units drive the generation in specific downstream tasks such as indirect object identification or concept promotion (Merullo et al., 2024; Geva et al., 2022). By examining activations immediately after these units, we aim to evaluate their respective contributions to emotion processing within each transformer layer. More formally, let $\mathbf{h}_t^{(l)} \in \mathbb{R}^d$ denote the hidden state vector at layer l and token index $t \in \{1, \cdots, T\}$, where d is the dimensionality of the model's hidden representations and T is the input sequence length. Then,

$$\mathbf{a}_t^{(l)} = \mathbf{MHSA}(\mathbf{h}_{1:t}^{(l-1)}),$$

$$\mathbf{m}_t^{(l)} = \mathbf{FFN}(\mathbf{h}_t^{(l)} + \mathbf{a}_t^{(l)}),$$

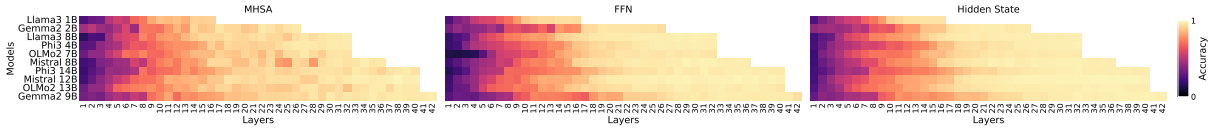

Figure 2: Layer-wise accuracies of emotion probe experiments across different models (each row) with varying depths at (**Left**) MHSA, (**Mid**) FFN, and (**Right**) hidden states. The results suggest an increasing signal with clear consolidation in the mid layers across various model families and sizes, which indicates that models predominantly make emotion-related decisions by the mid layers with minimal improvement in higher layers.

$$\mathbf{h}_t^{(l)} = \mathbf{h}_t^{(l-1)} + \mathbf{a}_t^{(l)} + \mathbf{m}_t^{(l)},$$

where $\mathbf{a}_t^{(l)} \in \mathbb{R}^d$ and $\mathbf{m}_t^{(l)} \in \mathbb{R}^d$ are MHSA and FFN outputs at layer l for token t. $\mathbf{h}_{1:t}^{(l-1)}$ are the previous layer's hidden states for tokens 1 to t.

Throughout this paper, we focus on activations $\mathbf{x}_t^{(l)}$ selected from one of the three candidates in $\{\mathbf{a}_t^{(l)}, \mathbf{m}_t^{(l)}, \mathbf{h}_t^{(l)}\}$ and study their properties at different layers and token positions. While activations can be extracted from any token index, we anticipate the strongest emotion signals to be present at the last token, as it directly influences the model's output prediction in our task setup. Therefore, when clear from context, we omit the subscript T while studying the last token. Also, we drop the superscript (l) when generally discussing any activation across different layers.

5 Probing for Emotion Signals

Building on the linear representation hypothesis (Mikolov et al., 2013b; Elhage et al., 2022; Park et al., 2024), we perform probing experiments to assess the presence and strength of emotion-related signals at different activations within the model. Specifically, we train linear classifiers (Hewitt and Liang, 2019) to predict the corresponding emotions. We formalize the linear classifiers as follows:

$$\hat{\mathbf{y}} = \mathbf{W}^\top \mathbf{x} + \mathbf{b},$$

where $\mathbf{x} \in \mathbb{R}^d$ denotes the activation vectors at one of the locations specified in the previous section. $\mathbf{W} \in \mathbb{R}^{d \times C}$ is the weight matrix for emotion classification, $\mathbf{b} \in \mathbb{R}^C$ is the bias vector, C represents the number of emotion classes, and $\hat{\mathbf{y}} \in \mathbb{R}^C$ denotes the predicted logits for each emotion class.

We perform probing separately over different activation locations and layers across the model for the last token index. The probing results in Figure 2, measured as the accuracy on a held-out test set, indicate that the models begin consolidating emotional information in the hidden states $\mathbf{h}^{(l)}$ neither too early nor too late, but predominantly around the mid-layers across all models. For example, in the first row corresponding to Llama 3.2 1B in Figure 2, the emotional signal peaks by layer $l = 10$ out of a total of 16 layers. Beyond layer 10, there is no significant increase in probe accuracy, suggesting that the model effectively captures emotional content by this stage.

There is no clear distinction in probing performance between $\mathbf{m}^{(l)}$ and $\mathbf{h}^{(l)}$. Measurements from FFN closely track the hidden state dynamics, showing a steady increase in emotional conceptualization that peaks around the mid-layers. However, the heatmap corresponding to $\mathbf{a}^{(l)}$ shows a more dispersed pattern while following the same consistent increasing trend observed in other locations.

Our experiments reveal that emotion-processing mechanisms in LLMs are most pronounced in the middle layers across model families and sizes. *These findings suggest that the model has largely determined the output emotion by the mid-layers, with subsequent layers adding little additional processing.* Our observation aligns with the understanding that higher transformer layers capture more abstract and task-specific features.

Lastly, to evaluate the hypothesis regarding the importance of the last token in causal modeling, we repeat the analysis on the last five tokens for Llama 3.2 1B in Appendix C.3. We observe a consistent increase in signal strength from earlier to later tokens, reinforcing the focus on the last token as the primary contributor to output generation.

6 Emotion Transfer by Activation Patching

Given evidence suggesting that the model's internal representation of emotional content stabilizes around the mid-layers, we explore causal intervention in these regions to test their functional importance. Specifically, we assess whether the output emotion of a *source* example can be transferred to a *target* example, with a different emotion, by selectively patching activations from the source com-

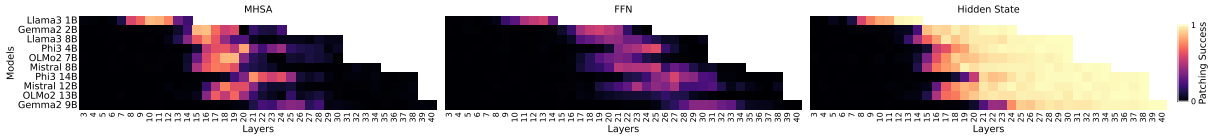

Figure 3: Results of activation patching experiments where we measure the success rate of transferring the output emotion by patching an activation from a source sample to a target sample. The patching is performed at (**Left**) MHSA, (**Mid**) FFN, and (**Right**) hidden states, respectively. The MHSA and FFN heatmaps demonstrate a clear localization with successful transfer peaks happening in the mid-layers consistently across various model families and scales. This observation aligns with the consolidation points observed in the probe heatmap (Figure 2) and indicates how activation patching can identify critical regions for emotion prediction.

putation graph into the inference pass of the target at corresponding locations, a method referred to as activation patching (Ghandeharioun et al., 2024).

Formally, let $\mathbf{x}_t^{(l)}$ be the activation vector from the target example, and $\hat{\mathbf{x}}_t^{(l)}$ be the activations from a different example, i.e. the source sentence, which has a mismatched label from the target. The patching operation involves replacing the activation at layer l and token t by substituting $\mathbf{x}_t^{(l)} \leftarrow \hat{\mathbf{x}}_t^{(l)}$ and letting the model continue the processing flow in the following layers and tokens.

The goal is to determine whether substituting specific activations with those from another example can manipulate the model's final prediction to reflect the intended label. We conduct experiments by substituting activations at the last token and within a window spanning five layers, consistently across all model families and sizes[1].

The results shown in Figure 3 demonstrate consistent behavior across model sizes and families. When intervening on hidden state activations, we observe a clear increase in intervention success (see Figure 3 right). Take Llama 3.2 1B as an example. Patching hidden states in the early layers is entirely ineffective. There is a critical point, e.g. layer 10 in this model, after which copying the hidden state transfers the emotion label with a high success rate. A high success rate at the final layers is naturally an expected behavior, as the residual stream of hidden states aggregates the information as it gets closer to the final layers. However, remarkably, the chance of success peaks and stabilizes around the mid-layers in all models, an observation that aligns with our findings in the previous section.

To dig deeper, we look at the patching effect of MHSA and FFN units (see Figure 3 left and middle, respectively), which provides clear evidence for functional localization of emotion processing. More precisely, interventions targeting $\mathbf{a}^{(l)}$ and $\mathbf{m}^{(l)}$ show clear evidence of success localized to specific mid-layers, e.g., MHSA units of layers $l \in [9-11]$ in Llama 3.2 1B. In other words, successful patching of both MHSA and FFN units occurs only in a subset of layers and predominantly in the middle rather than the final layers, with the FFN's most successful patchings happening only slightly later than the MHSA's patching. *We hypothesize that there are a few consecutive layers in each language model whose MHSA units are responsible for gathering emotional information from the rest of the tokens and integrating it into the hidden state of the last token. This mechanism is immediately followed by a processing in the subsequent FFN units.*

To complement these findings, we perform an additional experiment, where a set of activations is knocked out in the forward pass to assess their impact. The results in Appendix B consistently align with those of our activation patching and probing, reinforcing the evidence of functional localization in emotion processing. These observations hold across different models and experimental methods.

Our results in all previous sections are not prompt-dependent. Specifically, changes in the format, wording, structure, or the number of demonstrations in the prompt do not affect the findings (see Appendix C.5). Additionally, we provide a control experiment in Appendix C.4 to show that the identified units are not critical when performing a different yet syntactically similar task. In fact, the final layers are most critical for this isomorphic syntactic task, which differs significantly from the units we found for emotion processing.

To further explain our findings, we analyze the attention patterns in Llama 3.2 1B. Specifically, we record the top 3 tokens attended to by all attention heads in the last token of each layer. We conduct this analysis for all samples in the dataset, provid-

[1]Smaller models are more sensitive to fewer transferred layers, while larger models resist emotion transfer and require a larger span. For consistency, we use the same window size across all models.

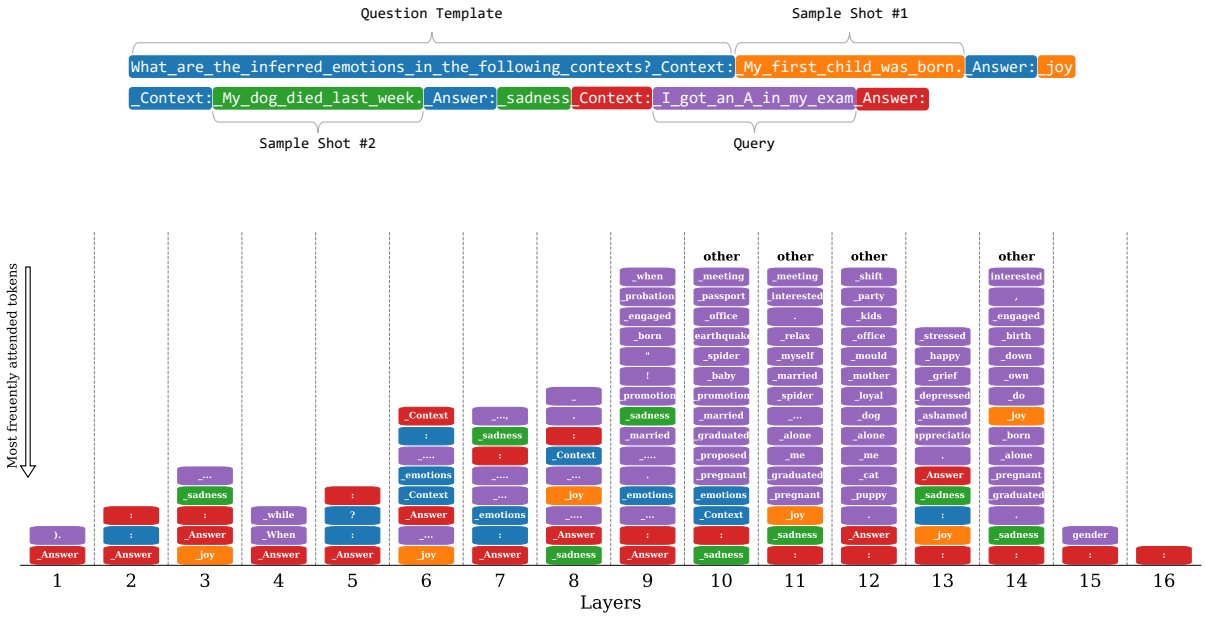

Figure 4: (**Top**) Primary prompt used in this study. Different segments of the prompt are colored differently. (**Bottom**) Most attended tokens at each layer from the perspective of the last token in Llama 3.2 1B. Layer 9 is the first layer in which the model attends to the tokens in the query with high emotional importance.

ing insight into which tokens are most frequently attended to. Figure 4 presents the aggregate attention patterns. The results indicate that early layers primarily focus on syntax. Around the mid-layers, the model shifts its attention to emotionally relevant tokens in the query and maintains this focus until the final layers. In the last few layers, attention predominantly focuses on the last token in the sequence, suggesting that it primarily carries forward the last token's hidden state from earlier layers. These patterns provide evidence that the identified middle units are meaningfully related to emotion processing.

7 Investigating Appraisal Concepts

We draw inspiration from cognitive appraisal theory to show an existing emotion structure in LLMs' latent representations. Appraisals are known to have significant associations with emotions and, in some accounts, are considered causal factors in their emergence (Rosenman and Smith, 2001).

Figure 5 illustrates three primary appraisal dimensions, including *pleasantness*, *other-agency*, and *predictability*, in addition to their associations with a set of basic emotions. For instance, *guilt* and *pride* are both materialized in situations with high *self-agency*, with the former happening in *pleasant* situations while the latter comes with an *unpleasant* experience. These mappings, which align closely with prior findings in appraisal theory (Wondra and Ellsworth, 2015), are extracted from our dataset by averaging appraisal scores for each emotion label.

We begin by training linear probes for appraisal signals within the model representations. In contrast to the linear classification probes used in Section 5, here we solve multiple independent regression tasks for each appraisal. More formally, consider a set of n appraisals and let $\mathbf{v}_a \in \mathbb{R}^d$ represent the weight vector corresponding to the appraisal $a \in \{1, \cdots, n\}$. Let $\mathbf{x}$ be an activation vector, as introduced in Section 4. We train the regression weight $\mathbf{v}_a$ and the bias $b_a \in \mathbb{R}$ such that,

$$\hat{r}_a = \mathbf{v}_a^\top \mathbf{x} + b_a,$$

where $\hat{r}_a \in \mathbb{R}$ is the estimate for the appraisal score a, given the input $\mathbf{x}$. We train a separate appraisal probe for each appraisal at the hidden states of each layer. The weight matrices obtained through this process serve as representations of the corresponding appraisal, encoding features of the appraisal concept at the activations of the specified locations. The success of appraisal probing highly depends on whether the target concept is linearly detectable in the targeted activation. We provide the appraisal probing results in Appendix E, showing that the appraisal signals are not linearly detectable at earlier layers but are strongly present as we approach the hidden state of the final layers.

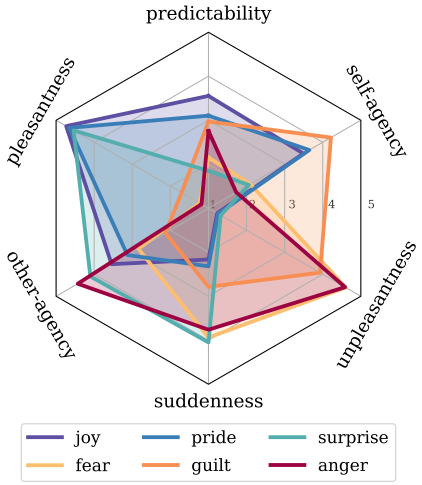

Figure 5: Appraisal emotion associations extracted from the dataset. For example, when participants report *anger*, they typically perceive a high degree of *other-agency* and a low level of *pleasantness* in the situation.

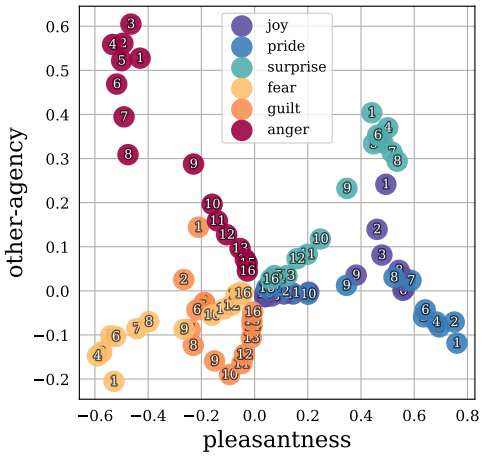

Figure 6: Cosine similarity of emotion vectors with *pleasantness* and *other-agency* appraisal vectors extracted from the hidden state of different Llama 3.2 1B layers. Layer number is written inside each marker.

8 Emotion-Appraisal Mappings

In this section, we analyze the representations of emotions and appraisals to reveal a structure within the latent LLM representations. Remember the weight matrix $\mathbf{W} \in \mathbb{R}^{d \times C}$ introduced in Section 5. Let $\mathbf{w}_e \in \mathbb{R}^d$ represent the column e of $\mathbf{W}$ corresponding to the emotion index $e \in \{1, \cdots, C\}$. We define the cosine similarity of appraisal a with emotion e as $sim(a,e) = \frac{\mathbf{v}_a^\top \mathbf{w}_e}{\|\mathbf{v}_a\|_2 \|\mathbf{w}_e\|_2}$.

Figure 6 shows the similarity score of emotion vectors with two appraisal vectors, i.e. the *pleasantness* and *other-agency*, throughout Llama 3.2 1B layers. Notably, we observe psychologically plausible appraisal-emotion mappings across all layers. However, the projection strength peaks in the early layers and fades to near zero in the final layer, suggesting orthogonality in the final layers.

We hypothesize that in the earlier layers, there exists a meaningful structure capturing appraisal and emotion concepts, which aligns with our expectations from the appraisal theory. However, these concepts gradually decouple as the processing progresses through the network, and by the final layers, they become orthogonal and fully decoupled, reflecting the specialization of the network toward higher-level tasks. We finish this section by drawing the connection to our findings in previous sections. Notice that the decoupling starts around the critical layers, e.g. layers 9 and 10 in Llama 3.2 1B, which we identified in previous sections. Therefore, our observations suggest that *the appraisals build a foundation to understand emotion representations in LLM hidden states, but the structure vanishes as we progress through the network.*

9 Intervention on Appraisal Concepts

After finding the appraisal vectors, we investigate the possibility of indirectly modifying the emotion of an input example by modulating its appraisals within the model representations. For this purpose, we need to isolate the role of each appraisal a, by considering its associated latent vector $\mathbf{v}_a$ and distinguishing it from other appraisal vectors. More precisely, we define $\mathbf{V}_{-a} := [\mathbf{v}_1, \ldots, \mathbf{v}_{a-1}, \mathbf{v}_{a+1}, \ldots, \mathbf{v}_n]$ by concatenating all appraisal vectors except $\mathbf{v}_a$. Next, we introduce the *unique effect vector* of appraisal a as $\mathbf{z}_a := (\mathbf{I} - \mathbf{P}_{-a})\mathbf{v}_a$, where $\mathbf{P}_{-a} = \mathbf{V}_{-a}(\mathbf{V}_{-a}^\top \mathbf{V}_{-a})^{-1}\mathbf{V}_{-a}^\top$ is the projection matrix onto the column space of $\mathbf{V}_{-a}$ and $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. We perform *appraisal modulation* by injecting $\mathbf{z}_a$ into the model's latent representation. More formally, the intervention is expressed as:

$$\mathbf{x} \leftarrow \mathbf{x} + \beta \frac{\mathbf{z}_a}{\|\mathbf{z}_a\|_2},$$

where $\mathbf{x}$ on the RHS is the original latent representation, e.g., a hidden state vector from a specific layer, and β is a scaling factor controlling the strength of the concept modulation. Notice that a positive β corresponds to an appraisal promotion while a negative β has the opposite effect of appraisal demotion. To measure the success of interventions, we evaluate the new emotion label

derived from this modification and repeat this procedure across all examples. Notice that, this modification is applicable to any layer of the model.

Figure 7 illustrates concept modulation results with different magnitudes of β on layer 9 of Llama 3.2 1B. Notably, we observe a remarkable alignment with theoretical and intuitive expectations. For instance, increasing the *pleasantness* appraisal promotes both *joy* and *pride*, aligning with the fact that both of these emotions have high associations with the *pleasantness* appraisal.

In contrast to these results, the appraisal modulation when applied to earlier layers, does not generate psychologically valid results and is totally ineffective when applied to later layers. This observation matches the intuitions on the mechanism we provided earlier; Intervening on the early layers is not valid since modifications to latent representations are overwritten by emotion-specialized mid-layers. On the other hand, intervention on final layers is not effective because of the orthogonality of concepts as demonstrated in Section 8. Appendix F presents the full results, including intervention across all layers of Llama 3.2 1B.

Appraisal theory is also predictive of the situations in which two appraisals are promoted simultaneously. To test this capability in LLMs, we perform an intervention on the superposition of two appraisal dimensions: *other-agency* and *pleasantness*, with mathematical details provided in Appendix F. The results, depicted in Figure 7, show a successful promotion of emotion *pride* with no further occurrences of *joy*. These findings provide strong evidence that layer 9 in Llama 3.2 1B directly contributes to cognitive processes related to emotions (see Appendix F for experiments with other appraisal concepts). Additionally, we provide complementary experiments such as direct emotion promotion using emotion vectors in Appendix D.

10 Discussion

We employed mechanistic interpretability techniques to investigate the inner workings of emotion inference in LLMs. Our results reveal that mid-layer MHSA units within these models are responsible for processing emotional content. By applying linear algebraic manipulations to modulate the antecedents of emotions, i.e. the appraisal concepts, we steered the model outputs in controlled and predictable ways. This is particularly important for ensuring the *reliability and steerability* of LLMs in high-stakes affective domains such as legal decision-making and clinical therapy.

A key distinction of our work is that we grounded our experiments on psychological theory and applied MI analysis on *in-the-wild examples*, rather than relying on *synthetically generated simplistic structures*, as seen in prior studies (Merullo et al., 2024). For example, Wang et al. (2023a) study the Indirect Object Identification task, by considering a fixed input structure, such as *"person1 and person2 had fun at school. person2 gave a ring to"* where the model is expected to predict *"person1"*. Hanna et al. (2023) study the "Greater-than" task using a dataset of examples like *"The war lasted from 1517 to 15"*, where the model is expected to predict any two-digit number larger than 17. While such notable efforts have demonstrated the feasibility of identifying end-to-end specialized circuits for these tasks, effectively interpreting them still requires structured inputs, making it challenging to generalize findings to more naturalistic settings.

Our work also highlights new opportunities for future research. Despite significant advancements in understanding human emotions, debates persist regarding the definition of emotion, the role of cognition in emotion, and the mechanisms underlying emotion inference (Ortony et al., 2022; Ellsworth and Scherer, 2003; Moors, 2013; Barrett, 2017). In parallel, cognitive neuroscience has explored the neural basis of emotion in support of differing theoretical perspectives (Kragel et al., 2024). The study of LLMs, combined with insights from emotion theory and neuroscience, opens a unique intersection for advancing our understanding of emotions (Sievers and Thornton, 2024).

Furthermore, our steering approach opens promising possibilities for conditioning LLMs to exhibit specific personality traits or moods, which could benefit applications requiring tailored affective responses (Jiang et al., 2023, 2024a; Petrov et al., 2024; Suh et al., 2024; Li et al., 2024; Suh et al., 2024). However, to ensure these interventions do not introduce unintended disruptions to other critical language-processing functions, it is essential to rigorously evaluate models on standard NLP benchmarks after inducing traits or moods.

Given LLMs' increasing societal impact —spanning areas such as mental health, legal decision-making, and human-AI interaction—it is imperative to deepen our understanding of their internal mechanisms. Our study breaks new ground in the

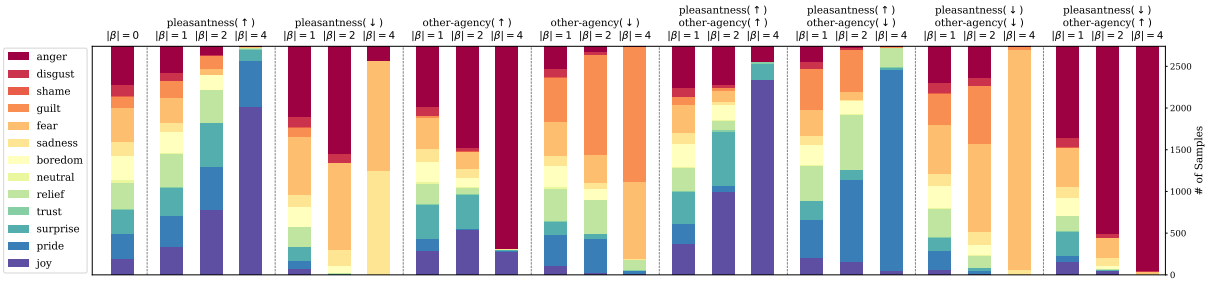

Figure 7: Results of appraisal concept modulation by intervening at layer 9 hidden states of Llama 3.2 1B for increasing scaling factors (β). $|\beta| = 0$ represents the original distribution of emotions in the dataset. For example, promoting (↑) or demoting (↓) *other-agency* significantly increases the share of *anger* and *guilt*, respectively. Similarly, promoting or demoting *pleasantness* increases the share of *joy*/*pride* and *sadness*/*guilt*/*anger* outputs, respectively. Additionally, promoting *pleasant other-agency* significantly increases the share of *joy* outputs, while the promotion of *unpleasant other-agency* significantly increases the share of *anger*.

interpretability of emotion inference in LLMs, offering a novel way to causally intervene in emotional text generation. These findings hold promise for improving safety and alignment in sensitive affective domains. Moving beyond black-box approaches to rigorously test and refine LLM emotion processing will not only advance the field of LLM interpretability but also unlock new pathways for more responsible AI systems.

11 Limitations

In this study, we build upon the linear representation hypothesis (Mikolov et al., 2013a,b; Levy and Goldberg, 2014; Elhage et al., 2022)—the idea that high-level concepts are encoded linearly within model representations (Park et al., 2024). This hypothesis is particularly appealing because, if true, it could enable simple and effective methods for interpreting and controlling LLMs—an approach we leveraged to localize and manipulate latent emotion representations. However, despite notable efforts to formalize the notions of linearity (Park et al., 2024) and orthogonality (Jiang et al., 2024b) in model representations, recent research suggests that not all features are encoded linearly (Engels et al., 2025). Further investigation is needed to improve clarity and robustness in this area.

Furthermore, we demonstrated the ability to manipulate affective outputs by modifying appraisal concepts. Nevertheless, the precise nature of this relationship remains unclear—it is possible that appraisals are merely correlated with emotions rather than exerting a direct causal influence or that the relationship follows an inverse causal pattern. Establishing causality requires further investigation in future studies to disentangle directional dependencies. A deeper understanding of the interplay between LLM emotional inference, emotion theory, and neuroscience will be crucial for both theoretical insights and practical applications. Addressing these challenges will refine our understanding of LLMs and enhance their reliability in affective computing.

12 Ethical Impact Statement

This study re-analyzes previously collected, de-identified data that had already undergone ethical review. The dataset is used for investigating the inner mechanisms by which auto-regressive LLMs process emotion. However, caution must be exercised when generalizing these results to models not examined in this work, to superficially similar tasks, or to different languages.

Our analysis highlights potential concerns for those deploying LLMs in high-stakes affective domains or for generating emotionally charged content. Given the risks associated with emotional manipulation by LLMs, it is crucial to develop a deeper understanding of how these models process emotions. To this end, we advocate for further research in this domain to ensure that LLMs align with ethical standards and human-centered AI principles.

Acknowledgments

This work is, in part, supported by the Army Research Office under Cooperative Agreement Number W911NF-25-2-0040. Only staff at ICT were sponsored directly by the Army Research Office. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, ei-

ther expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911, Albuquerque, New Mexico. Association for Computational Linguistics.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*.

Lisa Feldman Barrett. 2017. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23.

Yonatan Belinkov. 2018. *On internal language representations in deep learning: An analysis of machine translation and speech recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, Dublin, Ireland. Association for Computational Linguistics.

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

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Tobias Brosch and David Sander. 2013. Comment: the appraising brain: towards a neuro-cognitive model of appraisal processes in emotion. *Emotion Review*, 5(2):163–168.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. 2024. Black-box access is insufficient for rigorous ai audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition.

Phoebe C Ellsworth and Klaus R Scherer. 2003. *Appraisal processes in emotion*. Oxford University Press.

Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14:1199058.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are linear. In *The Thirteenth International Conference on Learning Representations*.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp*, pages 134–139.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. 2025. Monitoring latent world states in language models with propositional probes. In *The Thirteenth International Conference on Learning Representations*.

Nico H Frijda, Peter Kuipers, and Elisabeth Ter Schure. 1989. Relations among emotion, appraisal, and emotional action readiness. *Journal of personality and social psychology*, 57(2):212.

Chris Gagne and Peter Dayan. 2023. The inner sentiments of a thought. *arXiv preprint arXiv:2307.01784*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, Lisbon, Portugal. Association for Computational Linguistics.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Dieuwke Hupkes and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5617–5621. International Joint Conferences on Artificial Intelligence Organization.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *NeurIPS*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024a. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Yibo Jiang, Bryon Aragam, and Victor Veitch. 2024b. Uncovering meanings of embeddings via partial orthogonality. *Advances in Neural Information Processing Systems*, 36.

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

Philip A Kragel, David Sander, and Kevin S LaBar. 2024. Can brain data be used to arbitrate among emotion theories? In *Emotion theory: The Routledge comprehensive guide*, pages 511–542. Routledge.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. 2024. Shared functional specialization in transformer-based language models and the human brain. *Nature communications*, 15(1):5523.

Elisa Kwon, John D Patterson, Roger E Beaty, and Kosa Goucher-Lambert. 2024. Assessing the alignment between word representations in the brain and large language models. In *International Conference on-Design Computing and Cognition*, pages 207–223. Springer.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and S Yu Philip. 2024. Large language models in law: A survey. *AI Open*.

Richard S Lazarus. 1991. *Emotion and adaptation*. Oxford University Press on Demand.

Joana Leitão, Ben Meuleman, Dimitri Van De Ville, and Patrik Vuilleumier. 2020. Computational imaging during video game playing shows dynamic synchronization of cortical and subcortical networks of emotions. *PLoS biology*, 18(11):e3000900.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2024. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*.

Xin Liu, Ziyue Zhang, and Jingxin Nie. 2025. Talking to the brain: Using large language models as proxies to model brain semantic representation. *Preprint*, arXiv:2502.18725.

Ričards Marcinkevičs, Sonia Laguna, Moritz Vandenhirtz, and Julia E Vogt. 2024. Beyond concept bottleneck models: How to make black boxes intervenable?

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

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

MistralAI. 2024. Mistral nemo. Accessed: 2024-12-07.

Agnes Moors. 2013. On the causal role of appraisal in emotion. *Emotion Review*, 5(2):132–140.

Neel Nanda. 2022. Transformerlens.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Singapore. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Andrew Ortony, Gerald L Clore, and Allan Collins. 2022. *The cognitive structure of emotions*. Cambridge university press.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *International conference on machine learning*, ICML'24. JMLR.org.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

I Rosenman and C Smith. 2001. Appraisal theory: Overview, assumptions, varieties, controversies. *Appraisal processes in emotion. Theory, methods, research*, pages 3–19.

Klaus R Scherer et al. 1984. On the nature and function of emotion: A component process approach. *Approaches to emotion*, 2293(317):31.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Beau Sievers and Mark A Thornton. 2024. Deep social neuroscience: the promise and peril of using artificial neural networks to study the social brain. *Social Cognitive and Affective Neuroscience*, 19(1):nsae014.

Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Craig A Smith and Leslie D Kirby. 2011. The role of appraisal and emotion in coping and adaptation. *The handbook of stress science: Biology, psychology, and health*, pages 195–208.

Joseph Suh, Suhong Moon, Minwoo Kang, and David Chan. 2024. Rediscovering the latent dimensions of personality with large language models as trait descriptors. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Ala N. Tak and Jonathan Gratch. 2023. Is GPT a Computational Model of Emotion? . In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Los Alamitos, CA, USA. IEEE Computer Society.

Ala N. Tak and Jonathan Gratch. 2024. Gpt-4 emulates average-human emotional cognition from a third-person perspective.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, et al. 2024. Gemma 2: Improving open language models at a practical size.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*.

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.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023a. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023b. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Joshua Wondra and Phoebe Ellsworth. 2015. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological review*, 122.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. 2024. A reply to makelov et al. (2023)'s "interpretability illusion" arguments.

Nutchanon Yongsatianchot, Parisa Ghanad Torshizi, and Stacy Marsella. 2023. Investigating large language models' perception of emotion using appraisal theory. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

A Experimental Details

A.1 Dataset Details

For this project, we employed the crowd-enVENT dataset. Troiano et al. (2023) developed the dataset by asking crowdsource writers to share an event that made them feel a particular emotion. The vignettes varied in length, ranging from short sentences to longer narratives, but the language was predominantly everyday English. The participants were then asked to evaluate their subjective experiences during that event, including their perceived appraisals. Both were rated on a 5-point Likert scale with 5 representing the highest agreement.

The dataset is comprised of examples from the following list of emotions: *Joy, Pride, Surprise, Trust, Relief, Neutral, Boredom, Sadness, Fear, Guilt, Shame, Disgust, and Anger*. The dataset has 500 examples for each emotion, except for guilt and shame, which have 250 samples. The events were appraised along the following dimensions: *pleasantness, other-agency, predictability, suddenness, familiarity, unpleasantness, goal-relatedness, own responsibility, situational responsibility, goal support, consequence anticipation, urgency of response, own control, others' control, situational control, accepted control, internal standards, external norms, attention, not considered, and effort*. We selected a subset of dimensions for our analyses, previously shown to have a high association with emotions (Wondra and Ellsworth, 2015; Tak and Gratch, 2024).

To give a few examples from the dataset, the emotion label for the sentence "I baked a delicious strawberry cobbler" is *pride*, with appraisals *pleasantness* $= 5$, *other-agency* $= 1$, while "A housemate came at me with a knife" is an example of *fear* with *pleasantness* $= 1$ and *other-agency* $= 5$.

A.2 Architecture and Model Details

In this paper we experimented with ten LLMs: meta-llama/Llama-3.2-1B-Instruct and meta-llama/Llama-3.1-8B-Instruct (Grattafiori et al., 2024), google/gemma-2b-it and google/gemma-2-9b-it (Team et al., 2024), allenai/OLMo-7B-Instruct and allenai/OLMo-2-1124-13B-Instruct (OLMo et al., 2024), microsoft/Phi-3.5-mini-instruct and microsoft/Phi-3-medium-128k-instruct (Abdin et al., 2024), and mistralai/Ministral-8B-Instruct-2410 and nvidia/Mistral-NeMo-12B-Instruct (MistralAI, 2024). The architectural details for these language models are provided in Table 1. Unless stated otherwise, the default model used in this paper is Llama 3.2 1B since it is the lightest model, allowing for efficient analysis.

All models were implemented using the Hugging Face framework[2], leveraging the respective model weights, with additional integration of libraries such as TransformerLens (Nanda, 2022) to enable the hooking and intervention on hidden states and activations.

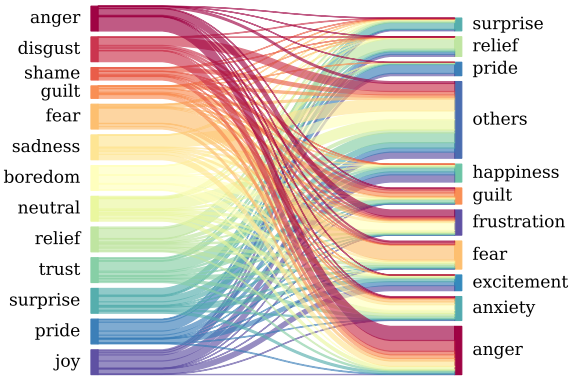

Figure 8: Llama 3.2 1B open vocab generation of emotions, comparing the true label to the predicted label through open vocab prediction

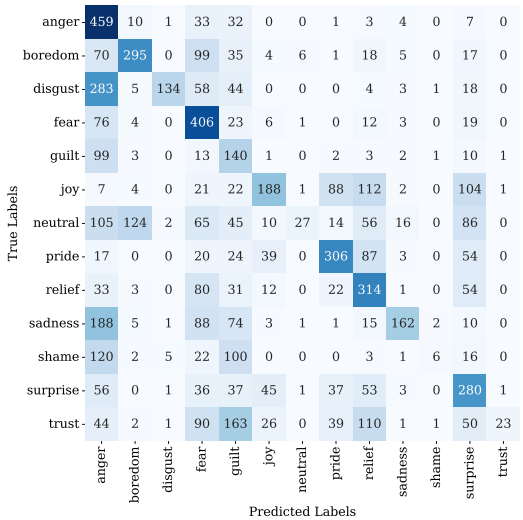

Figure 9: Confusion matrix comparing the true labels to Llama 3.2 1B predicted labels

A.3 Task Details

Emotion attribution is a challenging and subjective task, one that even humans often find difficult. The accuracy of third-person human annotations

[2] https://huggingface.co/models

Table 1: Architectural details of the language models used in this study.

	Llama 3.2 1B Instruct	Llama 3.1 8B Instruct	Gemma 2 2b-it	Gemma2 9b-it	Ministral 8B Instruct
Parameters	1B	8B	2B	9B	8B
hidden size d	2048	4096	2304	3584	4096
Layers	16	32	26	42	36
layer norm type	RMSNorm	RMSNorm	RMSNorm	RMSNorm	RMSNorm
Non-linearity	SiLU	SiLU	GeLU	GeLU	SiLU
Feedforward dim	8192	14336	9216	14336	12288
Head type	GQA	GQA	GQA	GQA	GQA
Num heads	32	32	8	16	32
Num KV heads	8	8	4	8	8
Context Window	131072	131072	8192	8192	32768
Vocab size	128256	128256	256000	256000	131072
Tied embedding	True	False	True	True	False

	Mistral 12B Nemo Instruct	Phi 3.5 mini Instruct	Phi 3 medium Instruct	OLMo 2 7B Instruct	OLMo 2 13B Instruct
Parameters	12.2B	3B	14B	7B	13B
hidden size d	5120	3072	5120	4096	5120
Layers	40	32	40	32	40
layer norm type	RMSNorm	RMSNorm	RMSNorm	RMSNorm	RMSNorm
Non-linearity	SiLU	SiLU	SiLU	SiLU	SiLU
Feedforward dim	14336	8192	17920	11008	13824
Head type	GQA	Multi-Head	GQA	Multi-Head	Multi-Head
Num heads	32	32	40	32	40
Num KV heads	8	32	10	32	40
Context Window	131072	131072	131072	4096	4096
Vocab size	131072	32064	32064	100352	100352
Tied embedding	False	False	False	False	False

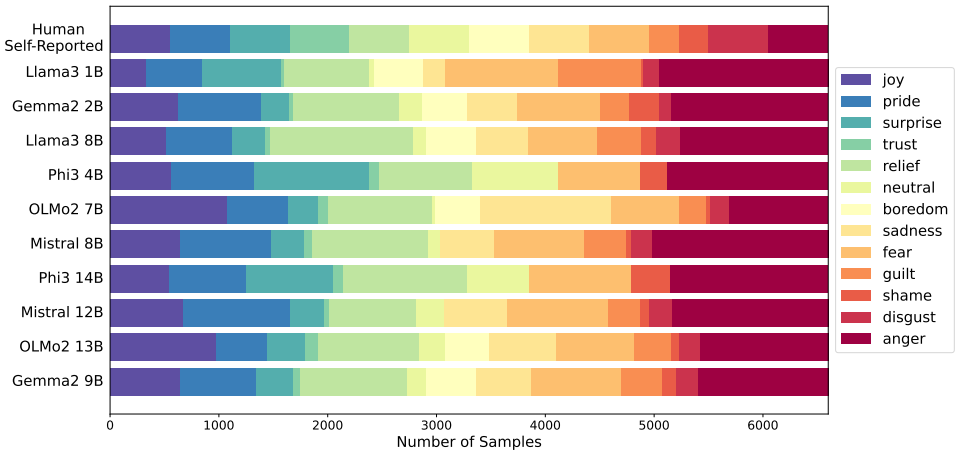

Figure 10: The distribution of next word emotion label predictions. The self-reported human labels are shown on the top row. The subsequent rows demonstrate the distribution of the predicted emotion labels from each model.

is approximately 50% in this dataset, as calculated using 1000 bootstrap resampling with 95% confidence intervals (the chance baseline being around 7%) (Troiano et al., 2023). Thus, it is not a surprise that defining a ground truth emotion label for each sample is challenging. Consequently, it is reasonable to expect that this task would also be challenging for LLMs. Figure 8 demonstrates the emotion label prediction on an open vocabulary, for Llama 3.2 1B. As the figure suggests, the LLM vocab choice in the output rarely matches the human-reported emotion label.

Therefore, we avoid doing an open-vocab generation, but instead, we confine the logits of LLM output to the set of emotion labels in the dataset. With this modification, we achieve an accuracy of approximately 40% or higher using self-reported ground truth emotions across all architectures and scales. Figure 9 demonstrates the closed vocabulary results through a confusion matrix for Llama 3.2 1B, comparing the true and predicted labels. Furthermore, Figure 10 illustrates the emotion label predictions across the LLMs tested, and Figure 11 demonstrates the accuracy results from these experiments.

For the rest of our analysis, we only focus on the correctly classified examples, which ensures at least 2,700 data points or more across different model architectures. We apply this filtering method to ensure that the samples selected for experimentation are ones where the LLM understood the emotion labeling task, allowing us to properly investigate the underlying mechanisms that led to the models' selected emotion label. However, we

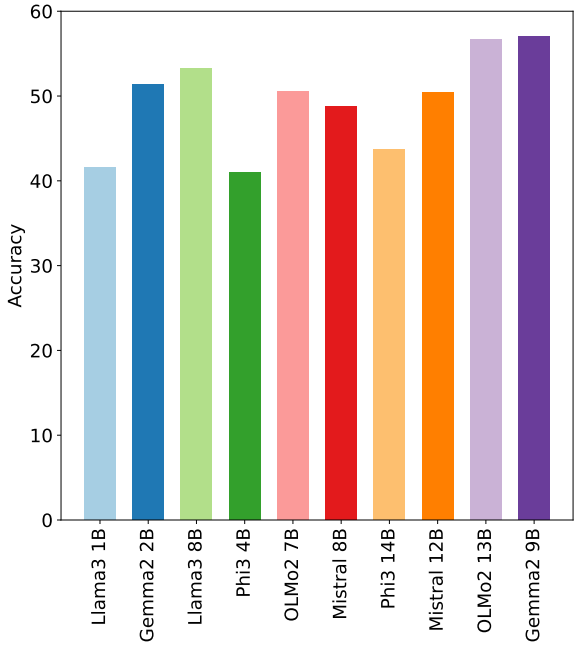

Figure 11: Comparison of models' accuracy on the emotion classification task. This experiment is performed across all samples in the dataset.

acknowledge that some emotion classes have a limited number of examples after filtering, which may present constraints in our experiments. A preliminary study targeting the bias of such filtering is provided in Appendix C.6, while deeper investigations are left for future studies.

Most experiments in this paper did not have a stochastic nature or behavior, and we used all the available data. The only source of stochasticity in our experiments comes from the train/test split in probe training. However, after extensive experiments, we noticed that the variance for these experiments is extremely low. More precisely, we observe no more than 1.5% STD in classification accuracy across 5 different seeds of each individual probe across all layers, with R^2 differing by no more than 0.02 between runs for regression tasks. The reported values in the paper figures are therefore only the mean values. Note that we used 5-fold cross-validation for probe hyperparameter tuning; therefore, no extra validation data split was necessary.

B Knockout Experiment

In this section, we elaborate on a causal approach in MI commonly referred to as *ablation*, *knockout*, or *zero/random activation intervention* (Rai et al., 2024; Olsson et al., 2022; Wang et al., 2023a; Chen et al., 2024). This is a complementary approach to the activation patching experiment to provide further evidence for the localization of emotion processing in LLMs. Precisely, we zero out the activations at different points in the model or replace them with random activations and assess the impact on the generated output label. In other words, we intervene in the activations in the forward pass of the model at MHSA, FFN, and the hidden states at the last token across different layers. Formally, let $\mathbf{x}$ be the activation vector from the target example at a specific layer and location. The zero-activation operation involves replacing the activation by substituting $\mathbf{x} \leftarrow \mathbf{0}$ and letting the model continue the processing flow in the following layers. Similarly, for random intervention, let $\mathbf{x}$ be the target activation to interrupt. The random activation intervention replaces $\mathbf{x}$ by substituting

$$\mathbf{x} \leftarrow \frac{\|\mathbf{x}\|}{\|\mathbf{r}\|}\mathbf{r},$$

where $\mathbf{r} \in \mathbb{R}^d$ is sampled from a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ distribution and the normalization factor $\frac{\|\mathbf{x}\|}{\|\mathbf{r}\|}$ ensures that the new activation has the same norm as the original one.

The modified activations propagate forward, affecting the model's outputs. Then, we compare the prediction coming out of the modified logits with the clean forward pass to measure the model's accuracy after the intervention. The lower this accuracy is, the more significant that activation is affecting emotion label generation.

In Figure 12, knockout-intervention across all models, we find remarkably consistent behavior with our probing and patching results provided in Sections 5 and 6—that after a certain point, even removing all MHSA units and, to some extent, the FFN units do not impact the final emotion classification accuracy. This indicates that the model's internal representation of the emotional content is established before that point. Additionally, we observe that knocking out activations with both zero and random interventions at $\mathbf{a}^{(10)}$ has a significantly greater impact than $\mathbf{m}^{(10)}$, which suggests that the MSHA unit in mid layers plays a more crucial role in collecting the emotion label from previous tokens.

For example, given the first row in the plot, we observe that zeroing out $\mathbf{a}^{(9-11)}$ and $\mathbf{m}^{(9-11)}$ LLama 3.2 1B has the greatest impact on the emotion label (corresponding to each example). As expected, activations at $\mathbf{h}^{(l)}$ have a significant impact

throughout all layers since they constitute the mainstream of the forward path reaching the model's output and are fundamentally different from $\mathbf{a}^{(l)}$ and $\mathbf{m}^{(l)}$, which contribute additional processing to the residual stream.

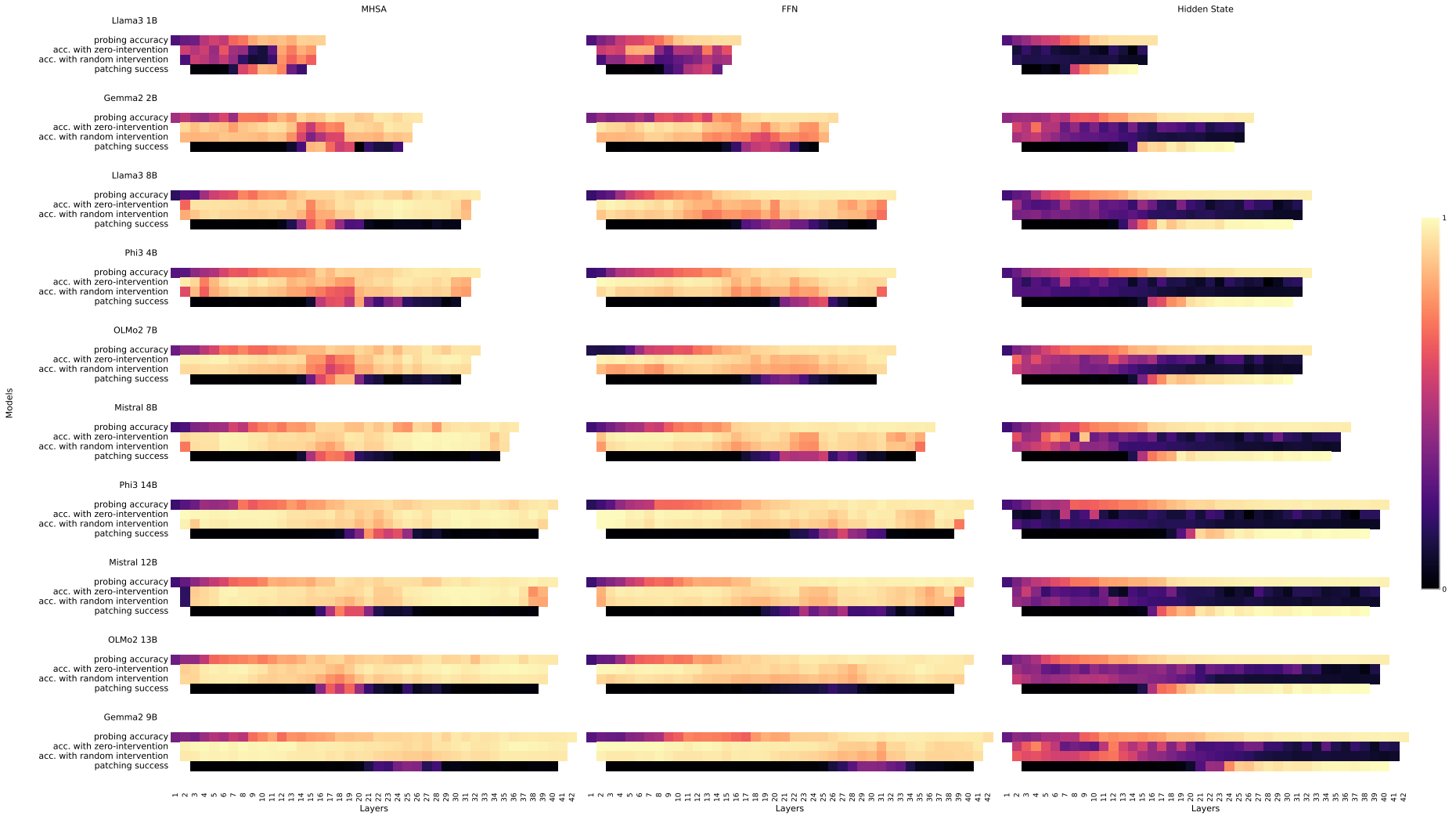

Figure 12: Comparison of probing, zero-activation and random-activation interventions, and activation patching on MSHA, FFN, and hidden state units across all layers of Llama 3.2 1B. The probe heatmap shows accuracy on the holdout set, zero/random activation interventions measure the model accuracy after disrupting causal pathways, and patching heatmaps indicate how effectively outputs transfer from source to target examples. The span sizes of 3 and 5 are used for the presented knockout interventions and patching experiments. This means that we intervene simultaneously on three/five consecutive layers, with the center being the indicated layer.

C Case Studies on Llama 3.2 1B

In this section, we focus on the Llama 3.2 1B language model and investigate the validity of our findings from multiple aspects. In Section C.1, we provide detailed results on emotion probing, activation patching and knockout interventions. In Section C.3 we extend our studies to include token dimension. Finally, we show that our results are robust to prompt design by conducting experiments using several hand-designed prompts in Section C.5.

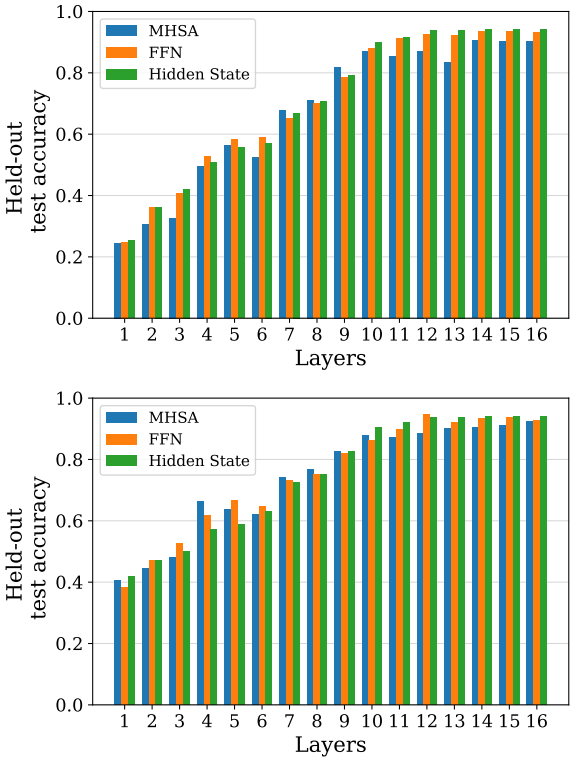

Figure 13: Probing test accuracy on last token of llama 3.2 1B for all layers. **Top** linear probe, **bottom** non-linear probe. There is a noticeable increase in probe performance in early layers when using a simple non-linear probe.

C.1 Details on Probing and Causal-Intervention

Here, we report the results of probing and causal intervention experiments on the Llama 3.2 1B model, focusing on the last token index across all layers, along with a simplified illustration of the findings. Figure 13 top demonstrates linear probing results.

It is noteworthy that we use linear probes to detect and extract emotion vectors. Low probe accuracy on earlier layers does not mean that there is no emotion signal at early layers but rather suggests that the signal is not *linearly* identifiable. In fact, Figure 13 bottom shows that when using a non-linear probe, i.e. a simple neural network with one hidden layer, the probe on earlier layers boosts considerably. However, both linear and non-linear probes peak around layer 10. After this layer, the model has finalized its output label decision and no major change happens.

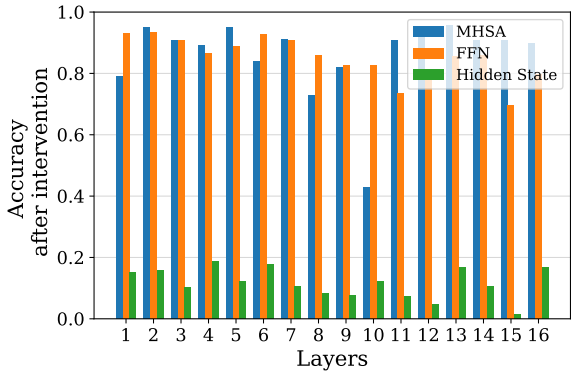

Figure 14: Zero-Intervention accuracy with span 1 on the last token index of Llama 3.2 1B across all layers. There is a clear drop in accuracy when MHSA activations in layer 10 are knocked out.

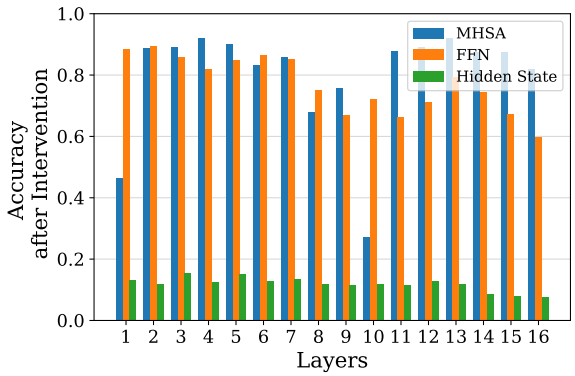

Figure 15: Random-Intervention accuracy with span 1 on the last token index of Llama 3.2 1B across all layers. A noticeable drop in accuracy is observed when MHSA activations in layer 10 are knocked out.

Intervention experiments further support this observation. Figures 14 and 15 show the effects of zero and random interventions with a span of 1, revealing a clear drop in accuracy when knocking out activations at layer 10. Finally, Figure 16 provides a detailed visualization of the patching experiment. The left-most plot highlights that the most successful emotion transfers occur at layer 10, which also exhibits the lowest number of unchanged labels. Notably, while some labels shifted to semantically similar emotions, they did not exactly match the

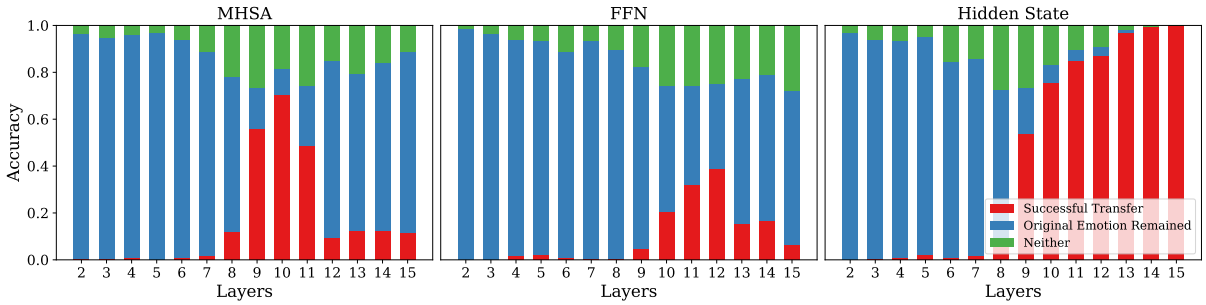

Figure 16: Activation patching results for Llama 3.2 1B across different layers (FFN, MHSA, and hidden state units) with Span = 3, evaluated over 200 source-target pairs. Blue indicates unsuccessful patching where the original label remained unchanged, red represents successful patching, and green denotes cases where the label changed but did not match the exact expected target.

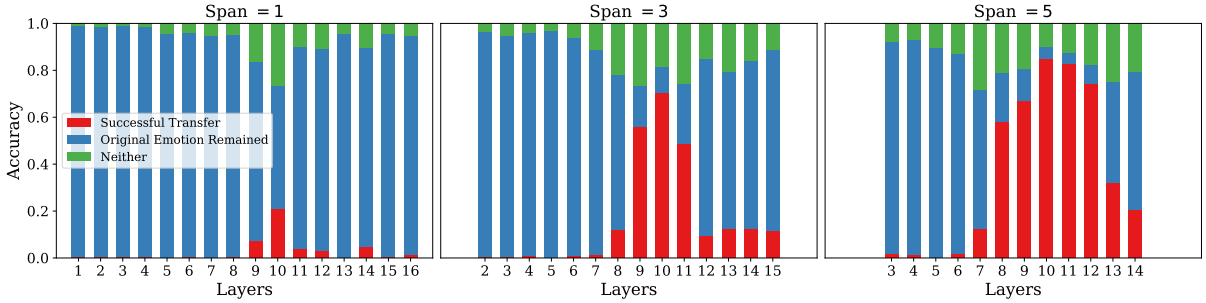

Figure 17: Effect of span size on Llama 3.2 1B activation patching at MHSA across different layers, evaluated over 200 source-target pairs.

target label and were, therefore, not counted as successful patches.

C.2 Effect of Span Size on Activation Patching

To examine the effect of span size on patching success, we repeated the experiment with three span sizes—1, 3, and 5 layers—on Llama 3.2 1B. Figure 17 presents the results, showing a clear increase in patching effectiveness as the span size increases. Notably, layer 10 continues to exhibit peak patching performance, reinforcing the idea of emotion-related functional localization in that layer.

C.3 Investigating Token Dimension

Throughout the paper, we extracted and analyzed the last token index activations $\mathbf{x}^{(l)}$—selected from one of the $\{\mathbf{a}^{(l)}, \mathbf{m}^{(l)}, \mathbf{h}^{(l)}\}$—across different layers of all tested models. We hypothesized that the strongest emotion signals appear at the last token, as it directly influences the model's next-word prediction in a causal language modeling setup. Earlier work demonstrates the presence of strong causal states immediately before the prediction, as well as their emergence at the final token of a noun phrase (Meng et al., 2022). Therefore, we suspect whether the last token of the query part in the prompt contains information more significant than the final token of the whole prompt.

To evaluate this hypothesis, we repeated the experiment on the last five tokens for Llama 3.2 1B extending the analysis to also include the final tokens in the query context (colored purple in Figure 4-Top). In Figure 18, we observe a consistent increase in signal strength from earlier to later tokens, reinforcing the focus on the last token as the primary contributor to output generation, but no clear importance on the last token of the query part.

Similarly, we conduct zero-activation intervention, random-activation intervention, and activation patching on Llama 3.2 1B's last five tokens across all layers. Figures 19, 20, and 21 confirm our hypothesis, suggesting that the last token's MSHA units in mid-layers, particularly $l = 10$, are critical for processing emotional content, while other token positions exhibit no localization.

Furthermore, the first row of Figures 19 and 20 illustrates the effect of zero or random activation interventions applied to all token positions in a layer

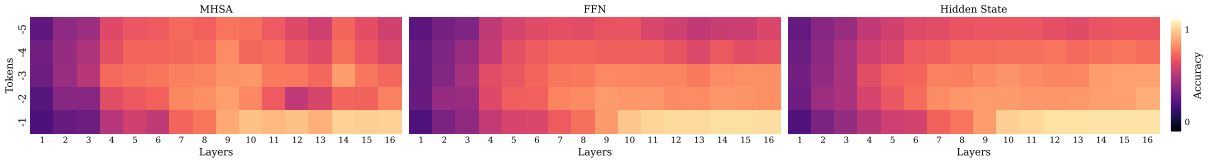

Figure 18: Probing test accuracy on different tokens of Llama 3.2 1B across all layers. We observe a consistent increase in signal strength from earlier to later tokens and from lower to higher layers.

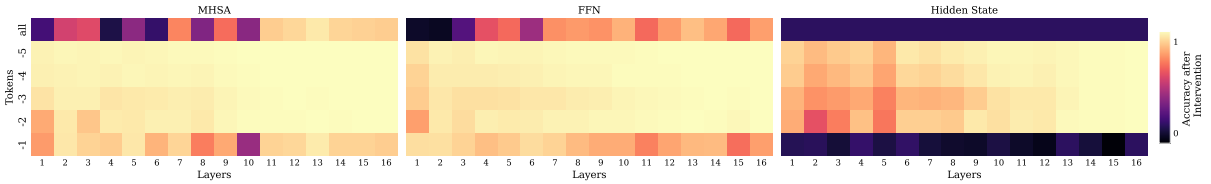

Figure 19: Zero intervention accuracy on different token indices of llama 3.2 1B for all layers with span = 1. The vertical dimension shows tokens and *"all"* means to knock out all activations of the specific layers at the same time. MHSA units beyond the critical layer ($l = 10$) have minimal impact on the causal path, even when an entire layer is knocked out.

(i.e., when all units in a layer are knocked out). Notably, the results indicate that MSHA units beyond the critical layer $l = 10$ contribute minimally to the causal path, even when an entire layer is knocked out.

C.4 Control Experiment on an Isomorphic Task

Here, we investigate whether the observed outcomes could be attributed solely to syntactic features and task structure rather than the target task of emotion processing. To assess this, we conduct an isomorphic experiment in which we modify the task to focus purely on syntax—predicting the first word in the sequence—and repeat the activation patching procedure as described in Section 6. The altered prompt is shown below:

> **What is the first word in the following contexts?** Context: My dog died last week. Answer: **My**; Context: I saw moldy food. Answer: **I**; Context: I could see my friend after a long time. Answer:

As shown in Figures 22 and 23, MHSA units in the final layers of the model are most critical for this syntactic task, which contrasts significantly with the emotion patching findings. Additionally, we observe weaker evidence of functional localization based on the patching results from the isomorphic control experiment.

C.5 Robustness to Prompt Design

To evaluate the model's robustness to various prompts, we designed a variety of prompt templates, as illustrated in Table 2. Figure 24 demonstrates the distribution of the next word emotion label predictions for the different prompt templates and different numbers of few-shot examples. Figure 25 demonstrates the final accuracy of these tests. Noteworthy that again in the following interpretability experiments, we confine our focus only on the samples that the model could predict correctly using each specific prompt.

Figures 26 and 27 validate that the probing and patching results presented earlier in the paper are robust to various prompt templates. For probing, we see that the model has determined the predicted label by the mid layers of the model, a result that is consistent across various prompt templates and the number of few shot examples provided. For activation patching, we also get consistent results across various prompt templates and with varied numbers of few-shot examples.

C.6 Investigating the Incorrectly Classified Samples

As mentioned earlier, we only used the dataset instances on which the LLM could predict the self-reported emotion label with no mistakes. We initially adopted this approach to reduce noise in the dataset, given the subjective nature of emotion labeling. Emotion classification can become a very subjective task in some contexts. For example, in many cases, it is hard to distinguish anger from

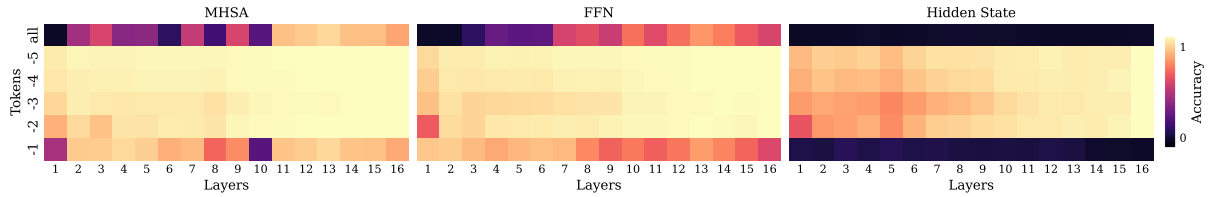

Figure 20: Random intervention accuracy on different token indices of Llama 3.2 1B across all layers with span = 1. The vertical axis represents token positions, where *"all"* denotes the simultaneous knockout of all activations in the specified layers. Notably, MHSA units beyond the critical layer ($l = 10$) contribute minimally to the causal path, even when an entire layer is deactivated.

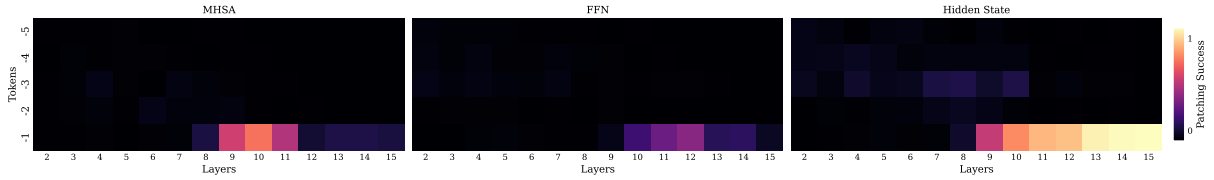

Figure 21: Results of activation patching success on different tokens of Llama 3.2 1B across all layers with span = 3. Clear functional localization is observed in layers 9–11.

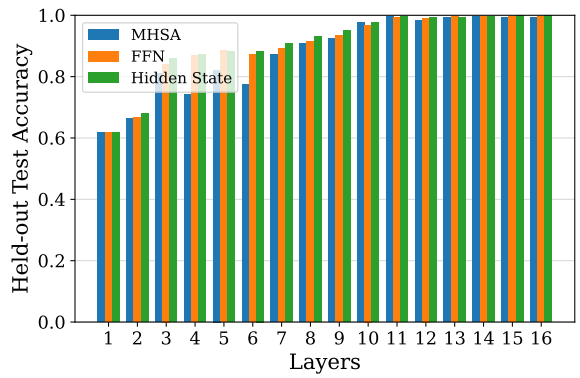

Figure 22: Probing test accuracy of the control isomorphic experiment on the last token of Llama 3.2 1B across all layers.

#	Prompt Template
1	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] Answer:
2	Consider this list of emotions: anger, boredom, disgust, fear, guilt, joy, pride, relief, sadness, shame, surprise, trust, neutral. 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] Answer:
3	Context: My first child was born. Answer: joy Context: My dog died last week. Answer: sadness Context: [Input] Answer:
4	Guess the emotion. Context: My first child was born. Answer: joy Context: My dog died last week. Answer: sadness Context: [Input] Answer:

Table 2: The prompt templates used for experimenting with the language models. The [Input] would be replaced with the sample sentence from the dataset that we are trying to label.

guilt, given the short description provided by the participants. Please see Figure 9, where we show how closely related emotions might be confused. We emphasize that focusing only on correctly classified examples may introduce bias, particularly a bias toward unambiguous emotional content. This is common in much of the prior work in mechanistic interpretability, which focuses on tasks that LLMs perform well on to ensure reliable interpretation.

That being said, analyzing misclassified examples could offer a deeper understanding of model mechanisms. In this section, we conduct a preliminary analysis comparing appraisal patterns between correctly and incorrectly classified cases for each emotion. Interestingly, we found that LLMs tend to exhibit distinct appraisal profiles in misclassified cases, suggesting shifts in the underlying reasoning process.

Table 3 summarizes these findings for three emotions that showed high confusion, presenting the mean and standard deviation of key appraisal dimensions using the appraisal probes we trained, along with t-test comparisons. These results suggest that misclassifications are not simply random errors but may stem from meaningful shifts in how

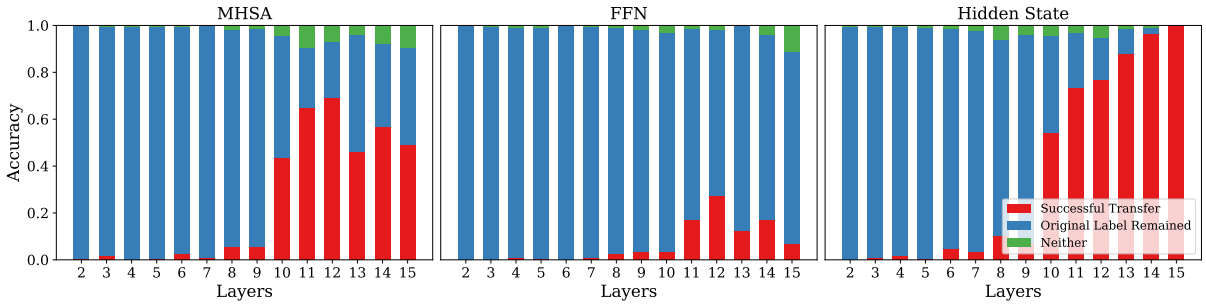

Figure 23: Control isomorphic experiment results for Llama 3.2 1B activation patching across different layers at MHSA, FFN, and hidden states with span = 3, evaluated over 200 source-target pairs. Localization is less evident, with the highest patching performance observed around the final layers.

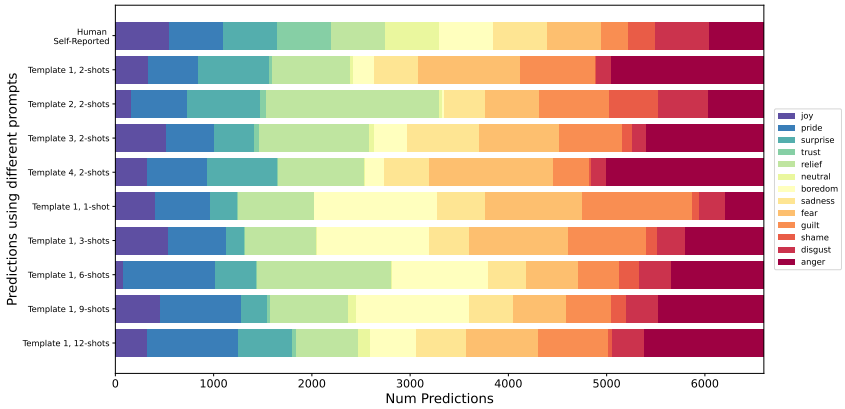

Figure 24: The distribution of next word emotion label predictions for different prompt templates and varied numbers of few shot examples.

the model evaluates appraisal cues.

Emotion	Pleasantness (Correct / Miss)	Others agency (Correct / Miss)
Sadness	$1.3 \pm 0.5 / 1.6 \pm 0.7$	$2.2 \pm 0.7 / 3.2 \pm 0.9$
Joy	$4.5 \pm 0.5 / 4.2 \pm 0.7$	$3.5 \pm 0.7 / 3.1 \pm 0.8$
Guilt	$1.9 \pm 0.6 / 1.6 \pm 0.8$	$2.8 \pm 0.7 / 3.3 \pm 0.7$

Table 3: Comparing the average appraisal values the LLM perceived for correctly classified vs misclassified samples with different emotion labels. All values reported here have p-value less than 0.001

D Direct Emotion Promotion

In Section 9, we showed that it is possible to change the model's output by manipulating appraisal concepts and directing it toward emotions with certain specifications of appraisals. In this section, we show that one can also directly inject a desired specific emotion label output by linearly adding the corresponding emotion vector to the hidden state of the model. More formally, recall the weight vector $\mathbf{w}_e$ for emotion e as introduced in Section 8. We define the emotion promotion modification as

$$\mathbf{x} \leftarrow \mathbf{x} + \beta \frac{\mathbf{w}_e}{\|\mathbf{w}_e\|_2},$$

where $\mathbf{x}$ on the RHS is the activation from the original model at any desired layer or location, and β is the scaling factor that controls the strength of emotion promotion.

Figure 28 shows the results of direct emotion promotion when performed on the hidden states across different layers of Llama 3.2 1B for different target emotion labels. As the figure suggests, direct emotion promotion is not effective when applied to the early layers, which completely aligns with our prior results. However, after layer 9, the success chance greatly improves, especially for large enough values of β. Again, this is a validation of our previous findings which shows that emotion concepts are linearly accessible and modifiable after the mid-layers. But before these layers, even a direct modification may fail since it will be overwritten later by the subsequent layers.

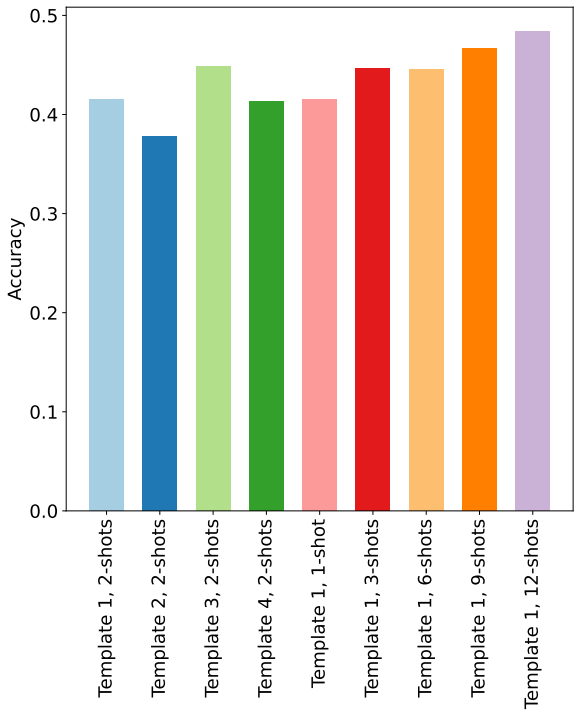

Figure 25: Accuracy of different prompts on the emotion classification task. This experiment varies both the prompt template and the number of provided few-shot examples. Experiments are conducted on Llama 3.2 1B.

E Appraisal Probing

As detailed in Section 7 and similar to emotion probing experiments provided in Section 5, we perform probing experiments to assess the presence and strength of appraisal signals at different activations within the Llama 3.2 1B model. We perform probing separately for each appraisal dimension over different activation locations and layers across the model for the last token. The probing results in Figure 29 are measured as the regression R^2 score on a held-out test set.

Following the behavior observed in probing emotion signals, models begin consolidating appraisal-related information in the hidden states $\mathbf{h}^{(l)}$ around the mid-layers. We observe that beyond layer 10, there is no significant increase in probe accuracy in any appraisal dimension. As also observed in the emotion probing experiment, there is no clear distinction in probing performance between $\mathbf{m}^{(l)}$ and $\mathbf{h}^{(l)}$.

As discussed in length in the main test, the success of linear probing highly depends on whether the target concept is linearly detectable given an activation. The results here also enforce the notion that the appraisal signals are not linearly detectable at earlier layers but are strongly present as we approach the hidden state of the final layers.

F Further Details on Appraisal Modulation

As discussed in Section 9, we show the possibility of indirectly modifying the emotion of an input example by modulating its appraisals within the model representations. In this section, we provide more details and further experiments on appraisal modulation.

First, we redefine our appraisal modulation method to generalize to cases where we want to modify multiple concepts simultaneously. Let's assume that we have a set of r appraisal vectors $\mathcal{A} := \{\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \cdots, \mathbf{v}_{i_r}\}$ which we want to modify, and consider a second set $\mathcal{B} := \{\mathbf{v}_{j_1}, \mathbf{v}_{j_2}, \cdots, \mathbf{v}_{j_k}\}$ to contain the k appraisal vectors which we want to maintain the corresponding appraisal concept fixed during the update. We define the project matrix $\mathbf{P}_\mathcal{B}$ as the projection matrix which projects into the span of $\mathcal{B}$. More formally, form the matrix $\mathbf{V}_\mathcal{B} = [\mathbf{v}_{j_1}; \mathbf{v}_{j_2}; \cdots; \mathbf{v}_{j_k}] \in \mathbb{R}^{d \times k}$ by concatenating the vectors in $\mathcal{B}$ as the columns of $\mathbf{V}_\mathcal{B}$. With this in mind, the projection matrix, $P_\mathcal{B}$ as

$$P_\mathcal{B} = \mathbf{V}_\mathcal{B}(\mathbf{V}_\mathcal{B}^\top \mathbf{V}_\mathcal{B})^{-1}\mathbf{V}_\mathcal{B}^\top,$$

and define the net effect vector as

$$\mathbf{z}_\mathcal{A} := (I - \mathbf{P}_\mathcal{B}) \sum_{a=i_1}^{i_r} \gamma_a \mathbf{v}_a,$$

where each γ_a is a variable from $\{-1, +1\}$ to indicate if the modulation promotes concept a or demote it. Finally, the modulation is performed as

$$\mathbf{x} \leftarrow \mathbf{x} + \beta \frac{\mathbf{z}_\mathcal{A}}{\|\mathbf{z}_\mathcal{A}\|_2}$$

where $\beta \in \mathbb{R}_+$ is a positive scaling factor.After performing the intervention, we measure the intervention's success by evaluating the new emotion label obtained by this modification across all examples in the dataset.

Section 9 reported the inference-time intervention results targeting the hidden state at layer 9 in Llama 3.2 1B, as we showed it is a critical point in emotion processing in Llama 3.2 1B. Here, we report the intervention results across all layers for a varied set of appraisal dimensions and their superposition to create more complex but specific concepts.

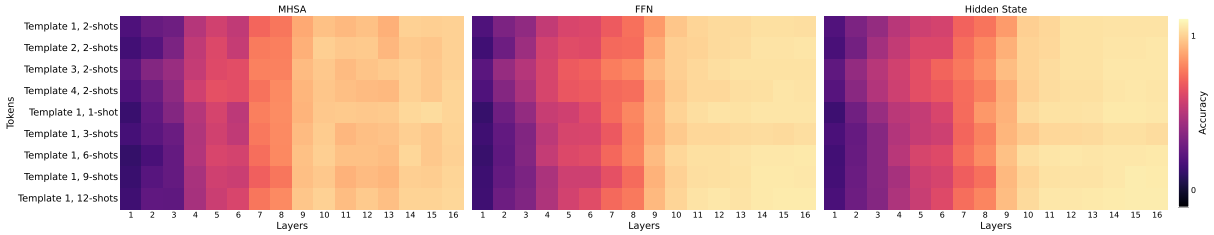

Figure 26: Probing accuracy with different prompts measured at last token in Llama 3.2 1B for all layers. This experiment varies both the prompt template and the number of provided few-shot examples.

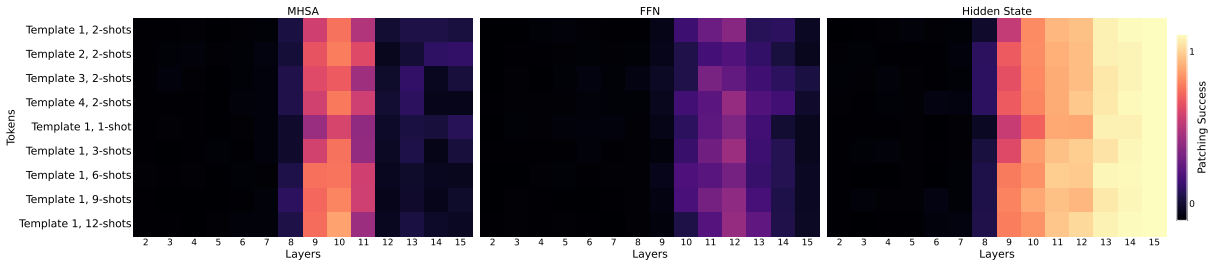

Figure 27: Success of activation patching with different prompts, measured at the last token index in Llama 3.2 1B across all layers with span = 3. This experiment varies both the prompt template and the number of provided few-shot examples.

As shown in Figures 30, 31, and 32, the intervention on appraisal concepts changes the distribution of the output labels and this distribution shift is intensified with higher values of β in all intervention experiments. However, the shift in the distribution of represented emotions does not necessarily conform with theoretical and intuitive expectations when intervening on earlier layers, particularly noticeable when β is sufficiently large. For example, in Figure 30, we note an unexpected decrease in the distribution of *joy* and *pride* in early layers, whereas psychologically plausible manipulations—such as an increase in high-valence emotions like *joy*, *pride*, and *surprise*, only emerge in mid-layers, peaking at layer 9. This observation supports the notion that intervening on the first layers is not effective because the linear structure in representations is not formed well yet.

Observing the intervention effect on later layers, we see significantly less pronounced distribution shifts. This also supports our earlier finding that the intervention on final layers is not effective because of the orthogonality of concepts that we showed in Section 8.

On the other hand, We observe a remarkable alignment with theoretical and intuitive expectations in the distribution shifts associated with interventions on middle layers (specifically layers 9-11). For instance, we observe that increasing the *pleasantness* appraisal promotes both *joy* and *pride*, aligning with the fact that both of these emotions have high associations with the appraisal. Also see Figure 35 for a different visualization.

To better evaluate intervention success, we also provide intervention results on the superposition of two appraisal dimensions (e.g., *other-agency* and *pleasantness*) across all layers in Llama 3.2 1B. The results, demonstrated in Figure 33, show a successful promotion of emotion *pride* with no further occurrences of *joy* in layers 9-11 when demoting *other-agency* and the promotion of *guilt* and *fear* with no occurrence of anger in mid-layers when demoting *other-agency*. Similarly, in Figure 34, we see the transition from *pride* to *surprise* and a larger distribution of *fear* as compared with *anger* when we promote *unpredictability*.

Finally, we will conclude this section by providing a control experiment in which, instead of using appraisal vectors, we intervene in the activation using a random vector. The results of this experiment are provided in Figure 36, and we see that no psychologically valid pattern is observable. Overall, these findings provide strong evidence that the mid-layers in Llama 3.2 1B meaningfully and directly contribute to cognitive processes related to emotions.

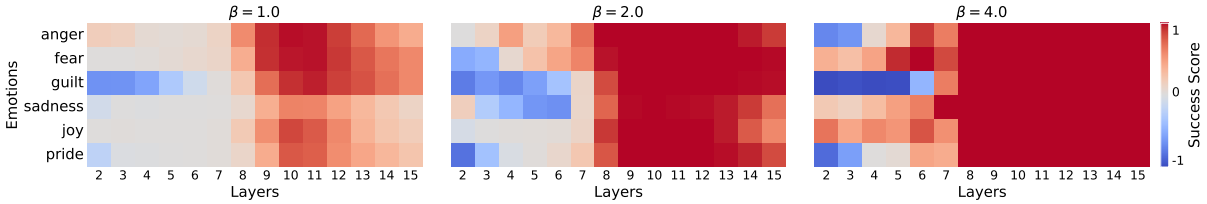

Figure 28: The heatmap showing the success of direct emotion promotion when applied at different layers of Llama 3.2 1B. Success score 1 means that the intervention successfully changed all output labels to the target emotion label. Scores 0 means that the intervention made no changes to the output and -1 means that the intervention resulted to complete opposite results, even damaging the samples with the correct original label. All the interventions in this figure used an intervention of layer span size 3.

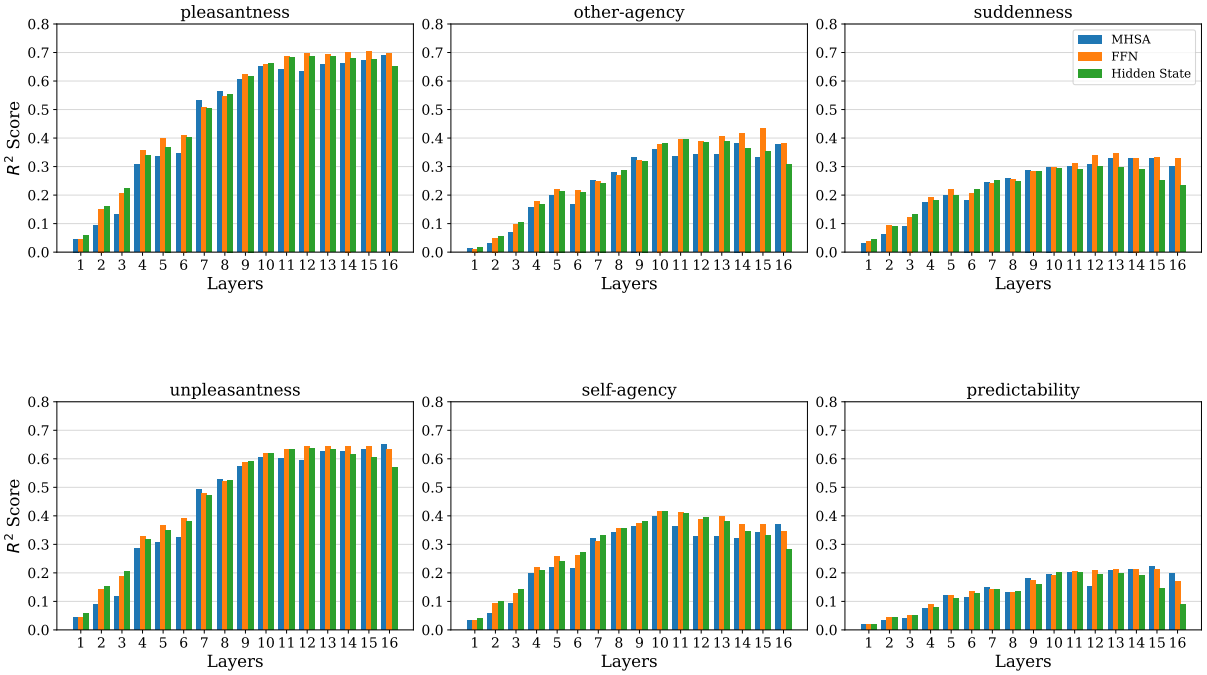

Figure 29: Probing results for Llama 3.2 1B, conducted separately for each appraisal dimension across different activation locations and layers for the last token. Results are measured as the regression R^2 score on a held-out test set.

G Code and Compute Resources

Our experiments are conducted using GPU-accelerated compute resources, with hardware such as NVIDIA A100 GPUs. For larger models, our studies are feasible on GPUs with at least 40GB of VRAM, with the full experiment running in approximately 24 hours. For smaller models, GPUs with 12GB of VRAM are sufficient to carry out our analyses efficiently.

Generative AI tools are utilized to improve the tone and style of writing, as well as for code completion during implementation.

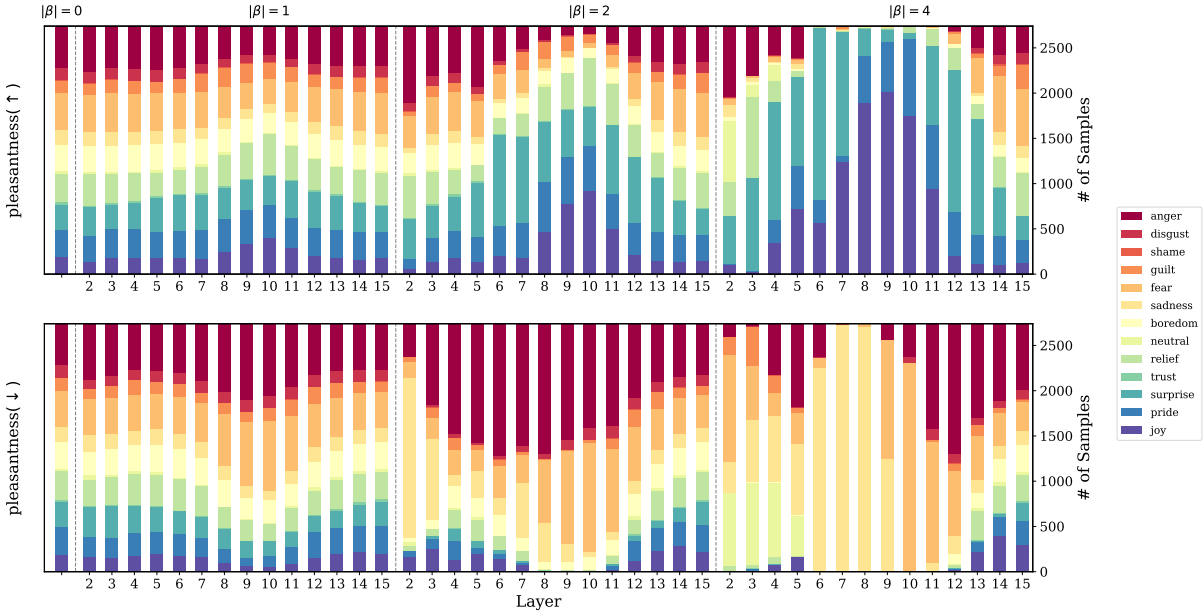

Figure 30: Effect of promoting and demoting *pleasantness* at different layers of Llama 3.2 1B with three levels of scaling factor β. $\beta = 0$ represents the original distribution without appraisal modulation. A consistent increase in distribution shift is observed as β increases across all intervention experiments. However, when intervening on earlier layers, particularly at higher β values, the shift in the distribution of represented emotions does not always align with theoretical expectations.

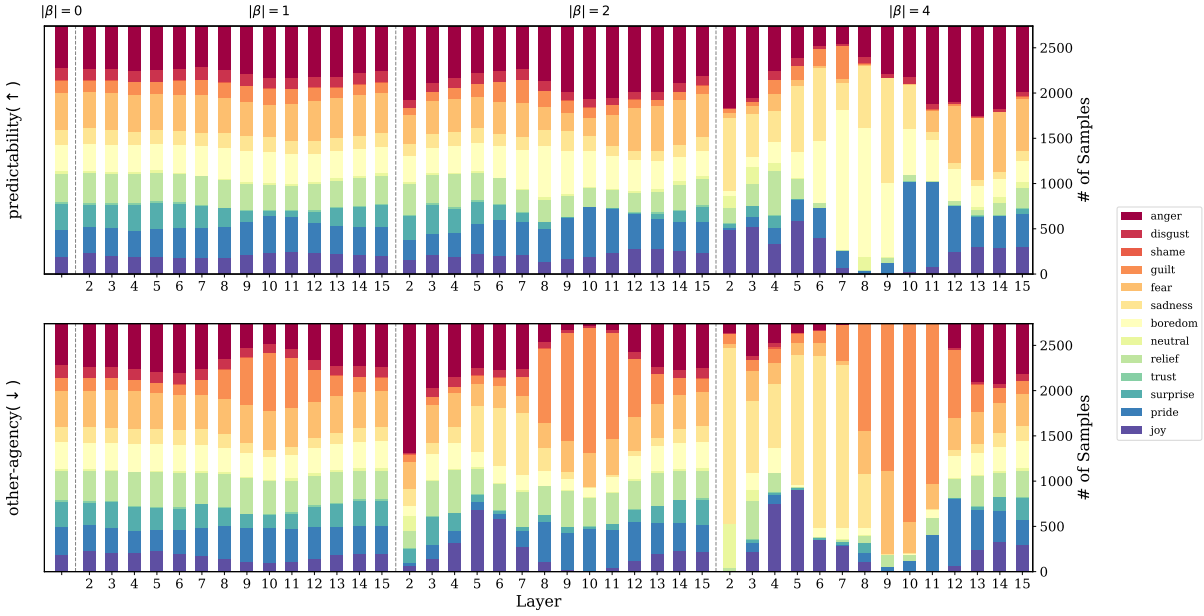

Figure 31: Effect of promoting and demoting *other-agency* at different layers of Llama 3.2 1B using three levels of scaling factor β. $\beta = 0$ represents the original distribution without appraisal modulation. Mid-layer appraisal modulation exhibits a theoretically plausible shift in emotion distribution.

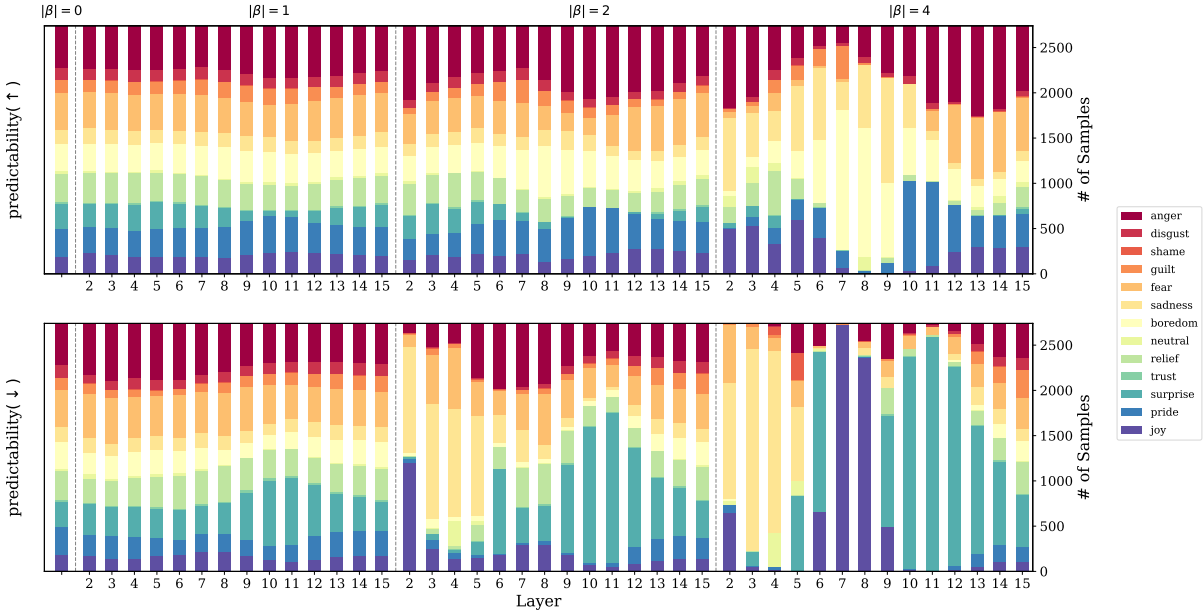

Figure 32: Effect of promoting and demoting *predictability* at different layers of Llama 3.2 1B using three levels of scaling factor β. $\beta = 0$ represents the original distribution without appraisal modulation. Mid-layer appraisal modulation exhibits a theoretically plausible shift in emotion distribution.

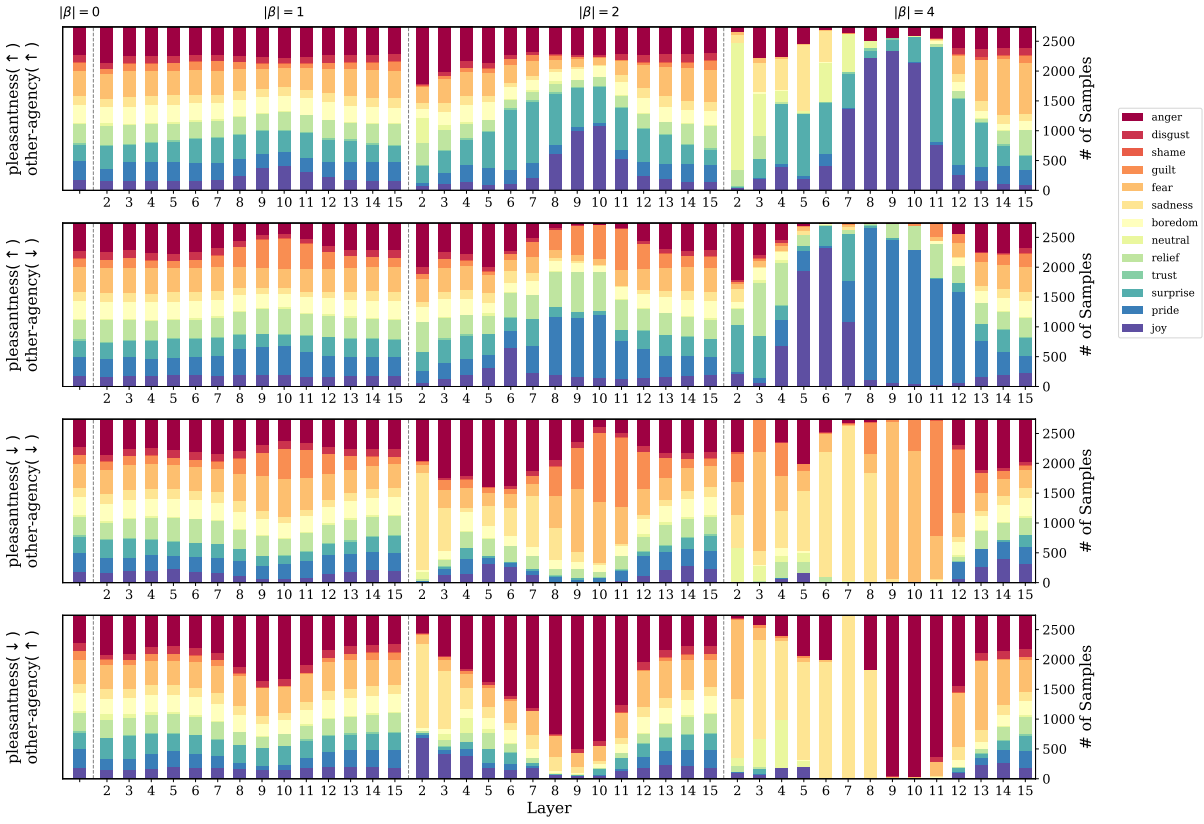

Figure 33: Superposition of *pleasantness* and *other-agency* appraisal modulation at different layers of Llama 3.2 1B. Results show successful promotion of *pride* with no further occurrences of *joy* in layers 9–11 when demoting *other-agency*, and the promotion of *guilt* and *fear* with no occurrences of *anger* in mid-layers when demoting *other-agency*.

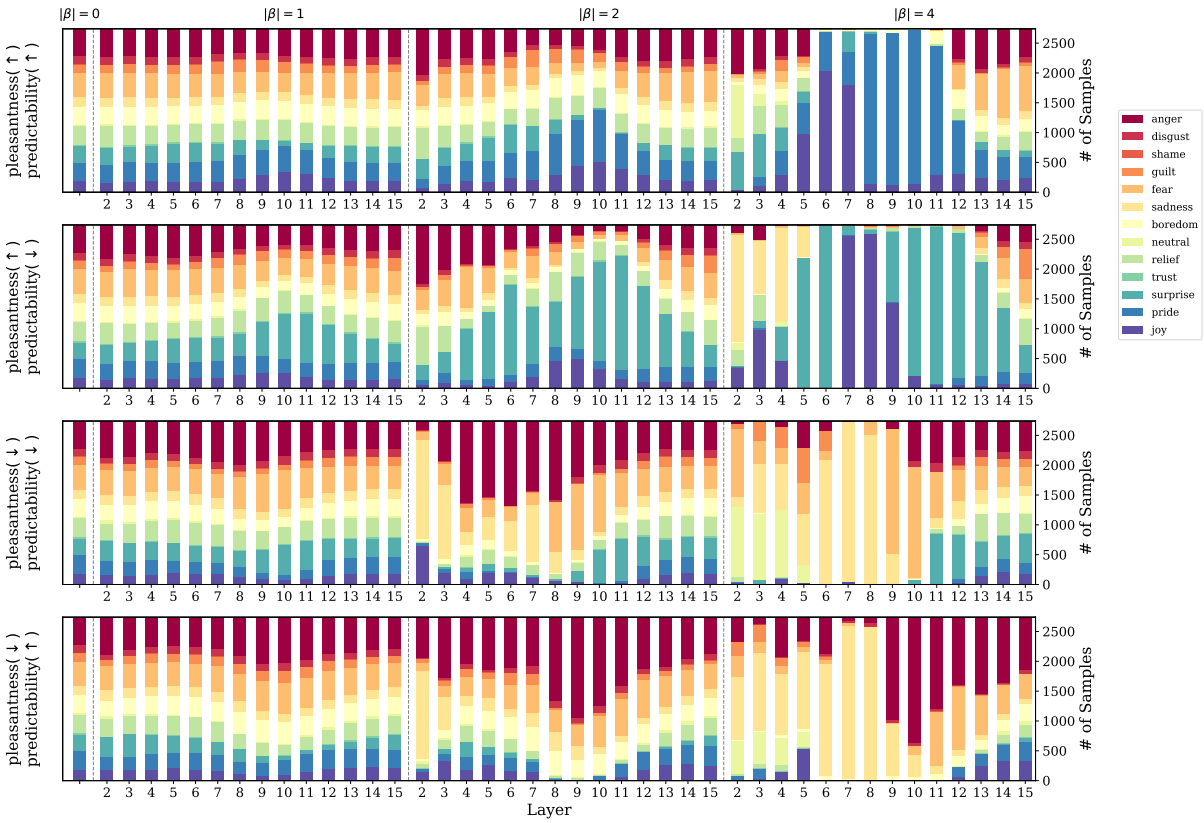

Figure 34: Superposition of *pleasantness* and *predictability* appraisal modulation at different layers of Llama 3.2 1B. Results show a successful transition from *pride* to surprise and a greater distribution of *fear* compared to anger in mid-layers when promoting unpredictability.

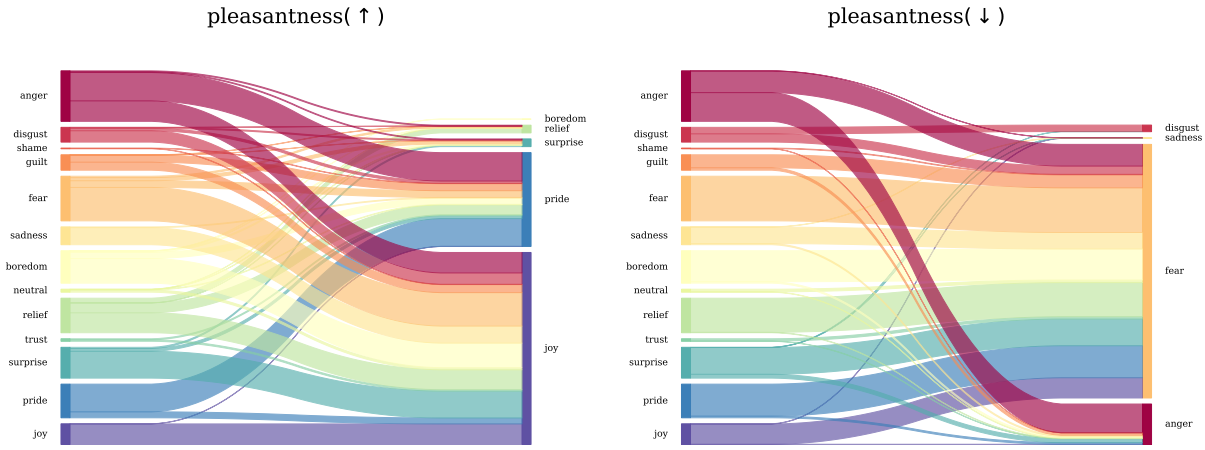

Figure 35: Sankey plot for *pleasantness* appraisal modulation when we perform it at layer 9 of Llama 3.2 1B model.

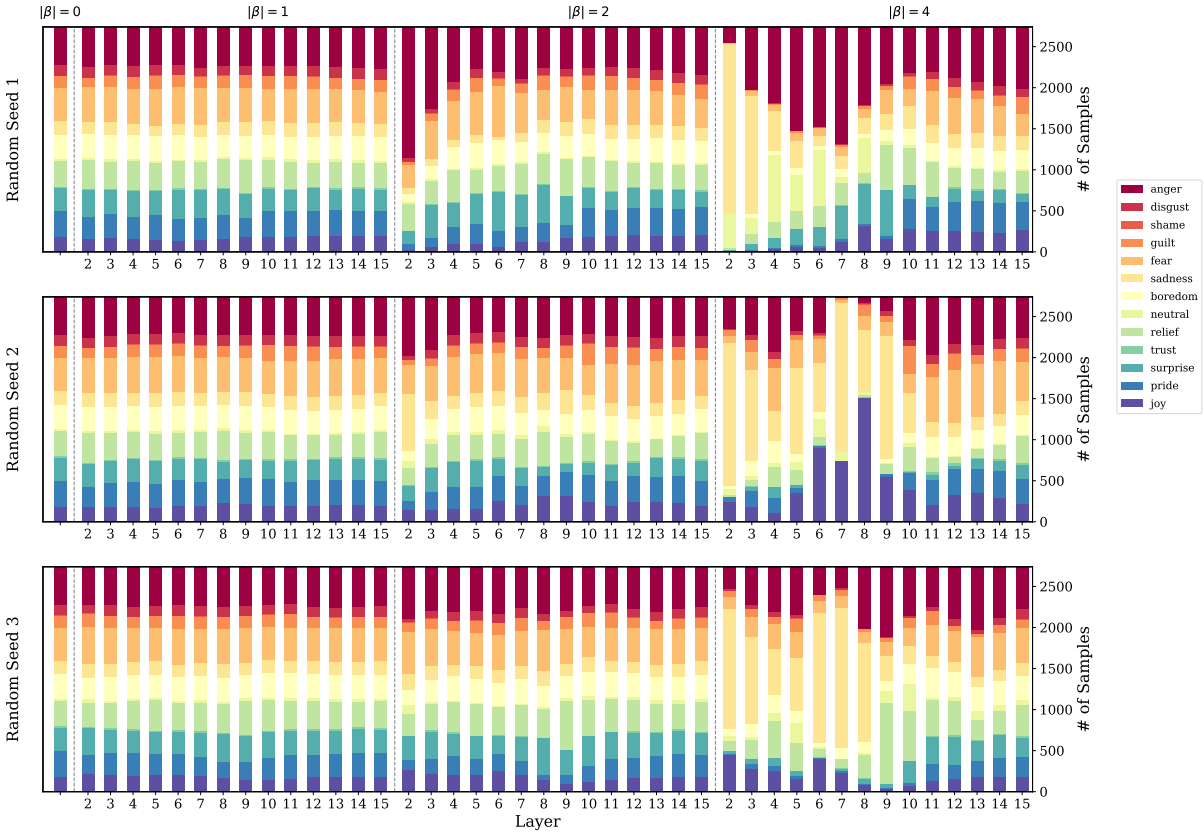

Figure 36: Results of the control experiment in which we randomly sample a vector and add it to the hidden state of the Llama 3.2 1B model at different layers. Each Row shows a different random seed.