FROM ENCODING TO GENERATION: TOWARDS EMPATHETIC LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Empathetic interaction involves two distinguishable capacities: representing another person's emotional state (cognitive empathy) and generating an appropriate response to that state (affective empathy). We apply this distinction to examine how emotion-related information is internally organized, selectively attended to, and carried through response generation in Llama-3.2-3B-Instruct across 10 emotion categories from the EmpatheticDialogues dataset. Layer-wise linear probing revealed a broad middle-layer region of peak emotion-category decodability, with the layer-11 L2 probe achieving 63.6% held-out accuracy against a 10% chance level and a 54.6% TF-IDF lexical baseline. During response generation, emotion-related user tokens received 1.89 times as much attention per token as other utterance tokens, concentrated across layers 10–15. Target-emotion alignment then decreased from a mean cosine similarity of 0.239 for input representations to 0.125 for generated responses, with positive drift in 77% of examples and the largest shifts for *angry* and *furious*. Emotion-category representation, attention allocation, and response-level alignment therefore emerge as distinguishable components of model processing that can be analyzed through a functional cognitive–affective empathy framework. These findings provide a foundation for monitoring emotion representations during generation and leveraging models' existing affective structure to support more naturally empathetic responses without relying on external prompting or explicit emotion conditioning.

1 INTRODUCTION

Empathy is a foundational capacity in human social interaction, encompassing two related but dissociable components. *Cognitive empathy* is the capacity to identify and encode what another person is feeling. *Affective empathy* is the capacity to generate a response that is appropriately aligned with or responsive to that perceived state. These components dissociate in humans; for example, a person may recognize others' distress without producing a supportive response or may respond emotionally without fully understanding the other's situation. Because empathy unfolds across both representation and response, studying it requires examining how emotional information is internally organized and how that organization shapes behavior.

Large language models (LLMs) are now routinely deployed in emotionally sensitive contexts like mental health support (Fitzpatrick et al., 2017), peer counseling (Sharma et al., 2021), and open-domain conversation (Rashkin et al., 2019). Despite producing outputs rated as highly empathetic by human evaluators (Lee et al., 2024), it is not known whether this behavior is grounded in structured internal representations of user emotions or arises from the surface statistics of empathetic language. The cognitive-affective empathy distinction provides a concrete framework for this investigation: does the model represent the user's emotional state, preferentially attend to emotionally relevant input, and carry that representation into its response?

1.1 RELATED WORK

Human empathy research distinguishes between representing others' emotional states and generating an appropriate response (Riess, 2017). Both components rely on distributed neural patterns rather than isolated emotion centers (Fan et al., 2011; Saarimäki et al., 2018), and structured emotion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

representations can be decoded from neural activity using multivariate pattern analysis (Baucom et al., 2012). The same decoding logic extends to transformer models: linear probes reveal what information is encoded at each layer, with different types of information emerging at different depths (Alain & Bengio, 2017; Belinkov, 2022).

LLMs trained with reinforcement learning from human feedback (RLHF) show a systematic positive-polarity bias, reliably generating positive emotional content while underserving negative emotions (Maheswaran et al., 2025). This suggests that empathetic-seeming outputs can arise from alignment-induced response patterns rather than from robust internal representations of user emotional states. Prior work has examined emotion information in LLM representations, but less is known about how such representations relate to attention allocation and the representational profile of generated responses.

Motivated by the cognitive-affective empathy distinction, we investigate Llama-3.2-3B-Instruct using the EmpatheticDialogues benchmark (Rashkin et al., 2019) to address three questions:

1. **Representation:** How is emotion-category information internally organized across the model's representational levels?
2. **Attention:** Does emotionally relevant input receive preferential attention during response generation?
3. **Response alignment:** How does emotion-related alignment change between the input representation and generated response, and how does this shift vary across emotion categories?

To address these questions, three analyses are combined: (1) layer-wise L1- and L2-regularized linear probing of all 29 representational levels to assess emotion-category encoding; (2) layer-wise attention analysis across token-source categories to assess preferential processing of emotional input; and (3) paired comparison of input and generated-response representations against emotion-category contrastive directions derived at the probe-selected layer. Together, these analyses reveal that emotion information is encoded in a broad middle-layer region (63.6% accuracy at layer 11), that emotionally relevant tokens receive preferential attention during generation (1.89 times enrichment), and that target-emotion alignment is only partially preserved in generated responses, with the largest decline for negative-valence emotions.

2 METHODOLOGY

This study examined how emotional information is represented and propagated in Llama-3.2-3B-Instruct (Grattafiori et al., 2024; Meta AI, 2024), which comprises 29 representational levels: one initial token-embedding layer and 28 transformer decoder layers. The study combined three analyses: layer-wise probing of hidden-state representations, attention allocation during response generation, and input–response emotion-trajectory analysis.

2.1 MATERIALS

The EmpatheticDialogues dataset (Rashkin et al., 2019), a crowdsourced collection of approximately 25,000 open-domain multi-turn conversations grounded in 32 emotion categories, was used. Ten categories were selected: *afraid*, *angry*, *anxious*, *devastated*, *lonely*, *sad*, *terrified*, *furious*, *grateful*, and *hopeful*. These span fear- and anxiety-related states (afraid, anxious, terrified), anger-related states (angry, furious), sadness and isolation (sad, devastated, lonely), and positive states (grateful, hopeful), enabling analysis of both fine-grained distinctions and broader affective contrasts.

For each of the 10 emotions, 400 non-overlapping situation descriptions were sampled to form Set 1, and a separate 200 first utterances were sampled to form Set 2. Set 1 was used for representation-level probing and prototype construction; Set 2 was used for response generation and generation-level analyses. Representative examples from each set are shown in Tables 1 and 2.

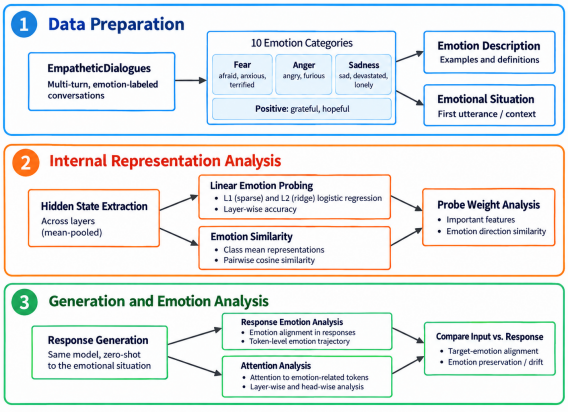

Figure 1: Overview of the analysis pipeline. Hidden states are extracted from emotion-labeled situations (Set 1) to probe and construct prototypes. First utterances (Set 2) are fed to the model in a zero-shot response-generation setting; attention and hidden-state trajectories are recorded during generation.

Emotion	Emotion description
Afraid	I was scared when I saw a mountain lion at night right in front of me. I ran to my tent.
Angry	Someone broke my window with a soccer ball.
Anxious	I drank all the coffee this morning and didn't share the whole pot with anyone. Now I feel very strange.
Devastated	Passing of my mother.
Lonely	My boyfriend went out of town for a few days, leaving me all alone at the apartment. I didn't have many friends here.
Sad	I just got laid off from my job. I need the paycheck, so this is not a good situation.
Terrified	A tornado came through my area last week
Furious	I was pissed when I saw someone put a dent in my car door.
Grateful	My mom let me move back in with her when my relationship didn't work out. I'm so thankful to have such a good mother.
Hopeful	I am going to the casino next weekend. I am entering a poker tournament and think I have a good chance.

Table 1: Examples of emotion-labeled descriptions from Set 1. Set 1 was used for hidden-state probing, prototype construction, and representational analyses. Text is reproduced from the dataset, with capitalization added only where necessary for consistent presentation.

2.2 Procedure

2.2.1 Hidden-State Extraction

Each text sample from Set 1 was passed through the model separately. Hidden-state representations were extracted from all 29 levels and averaged using masked mean pooling to produce a 3,072-dimensional sequence-level vector, excluding padding and the BOS token (Reimers & Gurevych, 2019; Di Palma et al., 2025). For each emotion and layer, an emotion prototype was computed by averaging the 400 pooled vectors belonging to that emotion. A grand-mean prototype was computed across all Set 1 examples; the target-emotion contrastive direction was defined as the difference between the emotion prototype and the grand mean.

2.2.2 RESPONSE GENERATION

From the first 200 utterances available for each emotion in Set 2, 20 were sampled for response generation, yielding 200 generated responses in total and introduced into the model's chat template using a zero-shot system prompt, positioning the model as a medical assistant instructed to produce a one- to two-sentence response (full prompt in Appendix A.1). Responses were generated autoregressively with nucleus sampling (temperature 0.7, top-p 0.9, maximum 50 new tokens). Hidden states and attention distributions were recorded at every generation step.

2.3 ANALYSIS

2.3.1 LINEAR EMOTION PROBING

To characterize how emotion-category information is organized across representational levels, Separate layer-wise multiclass logistic regression probes were trained. Each probe is trained independently at one of the 29 representational levels, following standard layer-wise probing methodology (Alain & Bengio, 2017; Belinkov, 2022). High probe accuracy at a given layer indicates that emotion-category information is linearly accessible from the representations at that depth, revealing where in the network emotion categories emerge.

Two regularization schemes were used to assess whether this information is sparse or distributed across hidden dimensions. **L1 (Lasso) regularization** produces sparse classifiers by driving many coefficients to exactly zero, testing whether accurate classification can be achieved using only a small subset of dimensions. **L2 (Ridge) regularization** penalizes large weights while retaining nonzero coefficients across many dimensions, allowing information distributed across the full hidden space to contribute (see Appendix A.2). Comparing the two schemes, therefore, reveals whether emotion categories are encoded in a few localized features or spread across many partially redundant dimensions.

Set 1 is split once into stratified training (80%) and test (20%) partitions with a fixed random seed. For each layer and regularization scheme, the penalty strength is selected by stratified 5-fold cross-validation on the training partition. The best layer is selected separately for L1 and L2 based on cross-validated accuracy, and each selected probe is evaluated once on the held-out test partition. We report held-out accuracy and macro-F1 as primary metrics. This procedure ensures that probe performance reflects generalizable accessibility of emotion information rather than overfitting to the training data.

2.3.2 TOKEN-GROUP ATTENTION ANALYSIS

To determine whether emotionally relevant input receives preferential processing during generation, we analyze how attention is allocated across token-source categories at each generation step. Attention distributions are aggregated over four mutually exclusive source groups: the BOS token, template tokens (system instructions and chat formatting), user-utterance tokens, and previously generated tokens. Emotion-related tokens are analyzed as a subset of the user-utterance group.

Emotion-related utterance tokens are identified using the contrastive directions derived from the Set 1 prototypes at layer 11, the layer at which linear emotion decoding peaked. A token is classified as emotion-related if the cosine similarity between its layer-11 hidden state and the target emotion's contrastive direction is at least 0.10. This links the attention analysis directly to the same representational space used in the probing analysis, so that emotion-related is defined by the model's own internal geometry rather than by surface lexical cues.

At each generation step, attention weights were summed within each source group. Because the groups differed in token count, mean attention per token was additionally computed for emotion-related and non-emotional utterance tokens. The per-token enrichment ratio was defined as the mean attention per emotion-related token divided by the mean attention per non-emotional utterance token. Its variation across layers was then examined to determine whether preferential attention occurred in the same middle-layer region in which emotion-category information was most decodable. This approach treats attention as an index of processing allocation, not as a causal explanation of output (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). It allows us to test whether the model actively prioritizes emotional content during response construction.

2.3.3 RESPONSE EMOTION TRAJECTORY

To measure how emotional information from the input is preserved or transformed in the generated response, we compare representations against the target emotion's contrastive direction at layer 11, the layer selected by the L2 probe. This direction is derived from the Set 1 emotion prototypes. For each emotion e, the prototype $\boldsymbol{\mu}_e^{(11)}$ is the mean of the 400 pooled layer-11 hidden states from Set 1 descriptions labeled with emotion e. The grand mean $\boldsymbol{\mu}^{(11)}$ is computed across all 4,000 Set 1 examples. The contrastive direction is the centered prototype vector:

$$\mathbf{d}_e^{(11)} = \boldsymbol{\mu}_e^{(11)} - \boldsymbol{\mu}^{(11)}. \tag{1}$$

This vector captures what distinguishes emotion e from the average emotion representation at the model's peak decoding layer. Cosine similarity was used to compare representations with this direction, making the measure invariant to vector magnitude.

For each generated response, three representations are extracted and compared:

1. **Pooled input utterance.** The layer-11 pooled hidden state of the user's first utterance, computed with the same masked mean-pooling procedure used to construct the prototypes. This represents the affective profile of the input as a whole.
2. **Pooled generated response.** The layer-11 pooled hidden state of the generated response text, computed after generation is complete. This represents the affective profile of the model's output as a whole.
3. **Token-level response states.** The layer-11 hidden state of each generated token, taken in its full generation context at the moment the token is produced. These states trace how target-emotion alignment evolves token by token as the response unfolds.

The pooled input and response representations are computed from raw text using the same pooling procedure applied to Set 1. The token-level states are computed in the full chat-template and generation context. Because these two kinds of representations differ in context and aggregation, their absolute cosine values are not directly comparable; we therefore analyze pooled and token-level trajectories separately. Target-emotion alignment is computed for each representation $\mathbf{h}$ as the cosine similarity to the target emotion's contrastive direction.

Emotion drift is defined as the difference between input alignment and response alignment:

$$\text{Drift}_e = \cos\left(\mathbf{h}_{\text{input}}, \mathbf{d}_e^{(11)}\right) - \cos\left(\mathbf{h}_{\text{response}}, \mathbf{d}_e^{(11)}\right) \tag{2}$$

Positive drift indicates that the generated response is less aligned with the target emotion than the input; negative drift indicates stronger alignment in the response. Drift thus quantifies the representational shift between the emotional content encoded in the input and the affectiveness produced in the output. Token-level trajectories are inspected to determine whether this shift arises gradually over the course of generation.

For a relative category-level measure, each input and response representation was also compared with all 10 emotion directions. The argmax prediction was the emotion whose direction produced the highest cosine similarity. Target-emotion argmax accuracy was the proportion of examples for which this predicted emotion matched the dataset label (see Appendix A.9).

3 RESULTS

3.1 EMOTION REPRESENTATION ACROSS MODEL DEPTH

Emotion-category information is linearly decodable from hidden representations at substantially above-chance levels. Cross-validation on the training partition selected layer 11 for the L2-regularized probe, which achieved a mean cross-validated accuracy of 65.6% and a held-out accuracy of 63.6% (macro-F1 = 0.635). The L1-regularized probe achieved 60.9% held-out accuracy. Both probes substantially exceed the 10% chance level and the 54.6% TF-IDF lexical baseline (Figure 2).

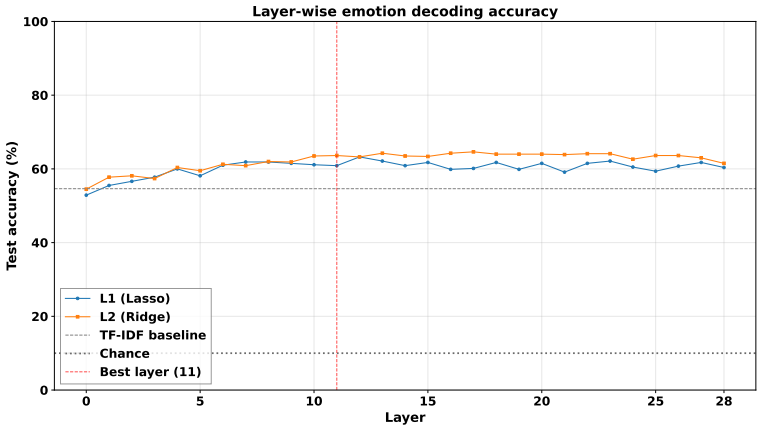

Figure 2: Test emotion-classification accuracy of L1- and L2-regularized probes across representational levels. Lines denote mean stratified five-fold cross-validation accuracy. Horizontal lines indicate the TF-IDF and chance baselines. The vertical marker indicates layer 11, selected by L2 cross-validation.

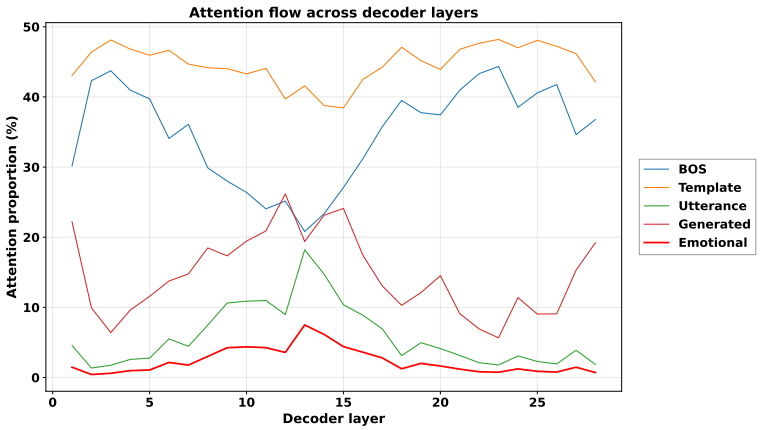

Figure 3: Attention allocation across decoder layers and token-source groups. Lines denote sample-level means across 200 generated responses.

Accuracy increased from the embedding layer through the early transformer layers, peaked at layer 11, remained broadly comparable across layers 8–17, and declined toward the final layers. Emotion-category information is organized across a broad middle-layer region, not localized at any single depth.

The dense L2 probe achieved 63.6% accuracy, compared with 60.9% for the sparse L1 probe. Emotion-category distinctions are distributed across many partially redundant dimensions rather than confined to a small subset of features. Pairwise separability, prototype geometry, and probe-weight analyses are reported in the Supplementary Material (Appendix A.7)

3.2 Attention Allocation During Response Generation

Averaged across responses, generation steps, layers, and attention heads, 44.7% of attention was directed to template tokens, 34.8% to the BOS token, 14.7% to previously generated tokens, and 5.8% to the user utterance. Within the utterance, emotion-related tokens received 1.89 times as much attention per token as non-emotional utterance tokens (Figure 3). Attention to emotion-related tokens varied across model depth. Peak allocation occurred at layer 13 (7.5%), with the next highest values at layers 14, 15, 10, and 11. This concentration across layers 10–15 descriptively overlaps with the broad middle-layer region of strongest emotion decodability.

3.3 Input-Response Emotion Alignment

The aggregate token-level trajectory showed lower target-emotion alignment at later response positions. Mean cosine similarity to the target-emotion direction was 0.239 for input utterances and 0.125 for generated responses, yielding a mean drift of 0.113. Positive drift was observed in 77% of the 200 paired examples. Drift varied strongly by emotion category (Figure 4). *Grateful* exhibited negative drift (-0.070), meaning the response was more aligned with the target emotion than the input, and *hopeful* showed near-zero drift (0.017). Positive drift increased across *lonely* (0.077), *sad* (0.085), *anxious* (0.086), *terrified* (0.135), *afraid* (0.140), *devastated* (0.166), *angry* (0.229), and *furious* (0.269). The two positive-valence categories showed the strongest retention of target-emotion alignment, whereas the anger-related categories showed the largest attenuation.

This pattern was confirmed by argmax classification: the target emotion was the highest-scoring direction emotion in 85% of *grateful* and 75% of *hopeful* response representations, compared with 0% for *devastated*, 10% for *furious*, and 20% for both *afraid* and *terrified*. Token-level trajectories further showed that target-emotion alignment generally decreased toward the end of each generated response (see Figure 7 in Appendix A.8)

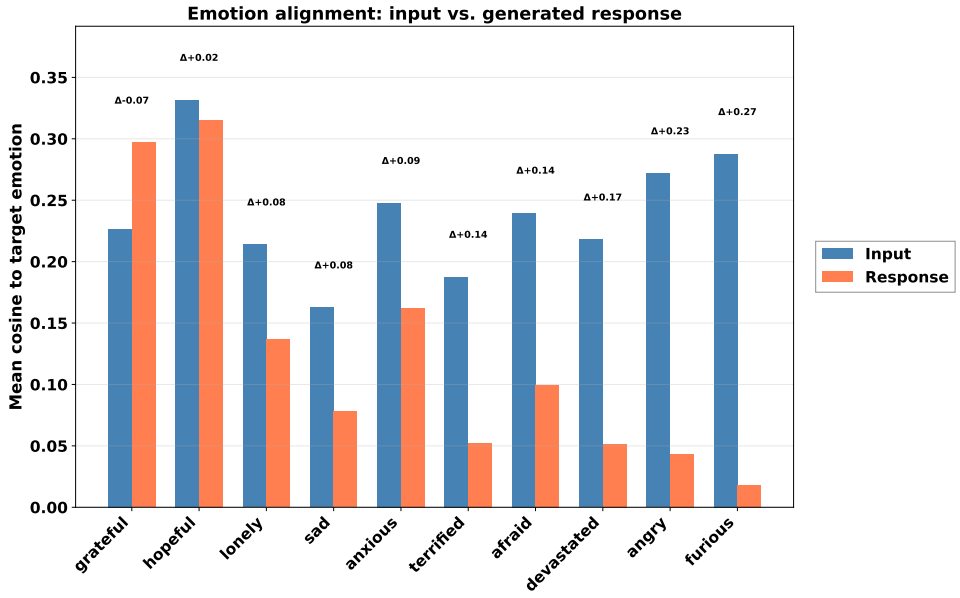

Figure 4: Target-emotion cosine alignment of input utterances and generated responses. Bars denote emotion-specific means across 20 paired examples. Emotions are ordered by mean drift (input alignment minus response alignment).

4 Discussion

The findings distinguish three components of emotion-related processing in Llama-3.2-3B-Instruct: emotion representation, attention to emotionally relevant input, and input–response alignment. Together, the results suggest that emotion-category information is internally accessible and selectively attended to during generation, but is not necessarily preserved in the model's response.

Emotion-category information was most accessible across a broad middle-layer region, with peak L2 probe performance at layer 11 and comparable performance across approximately layers 8–17. The sparse L1 probe retained 60.9% held-out accuracy after setting 63.4% of its coefficients to zero, compared with 63.6% for the dense L2 probe. This suggests that emotion-category information is distributed across multiple hidden features rather than localized to a single dimension. Pairwise analyses further showed that semantically distinct emotions were more separable than closely related categories such as *afraid* and *terrified*, indicating that the internal structure reflects relationships among emotion categories. This middle-layer accessibility is consistent with recent probing results

in Llama models (Di Palma et al., 2025) and, at a functional level, parallels evidence that human emotions are represented through distributed rather than isolated neural patterns (Lindquist et al., 2012; Saarimäki et al., 2018).

During generation, emotion-related utterance tokens received $1.89\times$ as much attention per token as non-emotional utterance tokens, with preferential attention concentrated across layers 10–15. The descriptive overlap between this region and the layers of strongest emotion decodability suggests a convergence between where emotion information is representationally accessible and where emotionally relevant content receives elevated attention during response construction. Thus, the model does not merely encode emotion-category information; it continues to preferentially allocate attention to emotion-relevant input as it generates a response. However, because attention reflects allocation rather than causal influence, intervention-based analyses are needed to establish whether these tokens or layers directly shape the response (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019).

Despite this representational and attentional sensitivity, target-emotion alignment decreased from input to response. Mean cosine similarity fell from 0.239 to 0.125, with 77% of examples showing positive drift. Importantly, this transformation was category dependent: alignment was retained or amplified for *grateful* and largely preserved for *hopeful*, whereas *angry* and *furious* showed the largest reductions. The argmax analysis showed the same tendency, with positive categories more frequently remaining the highest-scoring emotion direction, while several negative categories shifted toward other emotions. These results indicate that recognizing and attending to an emotional state does not imply that the same state will be preserved in the generated response.

This input-response shift may reflect two different processes. A supportive response may appropriately transform rather than reproduce the speaker's emotion; for example, responding to fear or anger with a calmer affective profile may be consistent with the listener's conversational role and with appraisal and regulation accounts of emotion (Lazarus, 1991; Riess, 2017). Alternatively, the shift may reflect alignment-induced preferences for positive or neutral language, particularly under the medical-assistant prompt.

The cognitive–affective empathy distinction provides a useful functional interpretation of these findings. Decodable emotion-category information provides a measurable analog of the representational component of cognitive empathy, while preferential attention indicates that emotion-relevant information remains salient during response construction. The subsequent input–response shift demonstrates that emotion representation and emotionally responsive generation are separable processes: the model can represent and attend to a speaker's emotion without simply reproducing it. This comparison is functional rather than experiential and does not imply human-like feeling or understanding. Instead, the framework decomposes emotion-aware generation into measurable stages: representation, attention, and response transformation, providing a basis for identifying where emotional information emerges, where it is prioritized, and how it changes as the model constructs a response.

4.1 Implications

Beyond characterizing emotion processing, these findings provide a mechanistic framework for evaluating emotion-aware generation in LLMs. By separating emotion representation, attention allocation, and response transformation, the framework can distinguish whether an emotionally inappropriate response arises from limited representation of the user's emotion, insufficient attention to emotion-relevant context, or a transformation that occurs during response generation. The overlap between emotion decodability and preferential attention in the middle layers further identifies candidate representations and model depths for future causal interventions. Rather than treating empathy as a property of the final response alone, these measures provide a basis for monitoring how emotional information is processed internally and for testing whether targeted interventions can improve the generation of emotionally appropriate responses without explicitly prescribing a particular emotion.

4.2 Limitations and Future Directions

The findings are correlational and scope-limited in several respects. First, linear probing demonstrates that emotion information is accessible from hidden states, but does not establish that these representations are causally used during generation (Hewitt & Liang, 2019; Ravichander et al., 2021). Similarly,

attention indicates where the model allocates processing focus rather than causal influence (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Future work could therefore use activation steering, ablation, or representation-engineering interventions to test whether the identified representations directly influence emotionally responsive generation (Zou et al., 2023).

Second, EmpatheticDialogues labels describe the broader emotional situation, and the target emotion may not always be explicit in the isolated first utterance used for generation. Future studies could validate emotion labels at the utterance level or provide the full situational context to better align the model input with the target emotion.

Third, the generation analysis used 20 responses per emotion under a single medical-assistant prompt, and the emotion set was weighted toward negative and closely related categories. Future work could use larger samples, varied conversational roles, and more balanced or dimensional emotion sets to assess the generalizability of the observed emotional drift.

The current analyses can identify when and how emotional representations change between input and response, but cannot determine whether such transformations produce more empathetic responses. Human evaluation and comparison with human listener responses will therefore be important for distinguishing contextually appropriate emotional regulation from a general tendency toward positive or neutral responses.

Finally, the study focuses exclusively on Llama-3.2-3B-Instruct. Applying the same framework across model families, model scales, and alignment procedures would determine whether the observed middle-layer organization and input–response transformation reflect broader properties of LLM emotion processing or characteristics specific to this model.

5 Conclusion

This study examined how emotion-related information moves from input encoding to response generation in Llama-3.2-3B-Instruct. Layer-wise probing showed that emotion categories were most accessible across a broad middle-layer region, with the layer-11 L2 probe achieving 63.6% held-out accuracy. During generation, emotion-related user tokens received $1.89\times$ as much attention per token as other utterance tokens, with the strongest allocation occurring in nearby middle layers. Despite this representational and attentional sensitivity, target-emotion alignment decreased from input to response in 77% of examples, with the largest shifts for *angry* and *furious*. Together, these findings distinguish three components of emotion processing in LLMs: representing a user's emotional state, attending to emotionally relevant content, and transforming that information during response generation. This framework provides a mechanistic basis for tracking emotion-related information across model depth and identifying candidate representations for future causal interventions aimed at improving emotionally appropriate generation.

6 AI Use Statement

Generative AI tools were used to assist with language editing, clarity, and polishing of the manuscript. The authors developed the research questions, methodology, analyses, interpretation of results, and scientific conclusions. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (Workshop Track)*, 2017. URL https://arxiv.org/abs/1610.01644.

Laura B. Baucom, Douglas H. Wedell, Jing Wang, David Blitzer, and Svetlana V. Shinkareva. Decoding the neural representation of affective states. *NeuroImage*, 59(1):718–727, 2012. doi: 10.1016/j.neuroimage.2011.07.037.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/COLI_a_00422.

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. LLaMAs have feelings too: Unveiling sentiment and emotion representations in LLaMA models through probing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6124–6142, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.306.

Yan Fan, Niall W. Duncan, Moritz de Greck, and Georg Northoff. Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience & Biobehavioral Reviews*, 35 (3):903–911, 2011. doi: 10.1016/j.neubiorev.2010.10.009.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2):e19, 2017. doi: 10.2196/mental.7785.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Kelsey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

John Hewitt and Percy Liang. 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)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357.

Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, New York, NY, 1991.

Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. Large language models produce responses perceived to be empathic. In *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 63–71, Glasgow, United Kingdom, 2024. IEEE. URL https://arxiv.org/abs/2403.18148.

Kristen A. Lindquist, Tor D. Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, 35(3):121–143, 2012. doi: 10.1017/S0140525X11000446.

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

Meta AI. Llama 3.2 model card. GitHub, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md. Released October 2024.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.

Helen Riess. The science of empathy. *Journal of Patient Experience*, 4(2):74–77, 2017. doi: 10.1177/2374373517699267.

Heini Saarimäki, Lara Farzaneh Ejtehadian, Enrico Glerean, Iiro P. Jääskeläinen, Patrik Vuilleumier, Mikko Sams, and Lauri Nummenmaa. Distributed affective space represents multiple emotion categories across the human brain. *Social Cognitive and Affective Neuroscience*, 13(5):471–482, 2018. doi: 10.1093/scan/nsy018.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021 (WWW '21)*, pp. 1817–1829. ACM, 2021. doi: 10.1145/3442381.3450097.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. 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)*, pp. 11–20, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A SUPPLEMENTARY MATERIAL

Emotion	First utterance
Afraid	I have a solo trip coming up. I'm going to Colorado to climb Mt. Elbert Solo. It's the tallest mountain in the United States.
Angry	On my way to work, a goose stole the toupee I needed to impress the new secretary, Christy.
Anxious	I have an important exam tomorrow and feel that something bad will happen.
Devastated	I had the worst night last night. I found out my uncle hit my cat with his car. It happened two weeks ago, and he never told me!
Lonely	I'm an over-the-road trucker. Man, it gets boring in the evenings.
Sad	Hello, yesterday was my birthday and my wife completely forgot!
Terrified	I moved into this new house and was woken up with this horrible screeching noise in the middle of the night.
Furious	I spent 5 hours baking a cake yesterday.
Grateful	My car broke down on my way to work yesterday.
Hopeful	I have been talking to this girl that I like for awhile now. I think that I am finally going to ask her on a date

Table 2: Examples of emotion-labeled first utterances from Set 2. First utterances were used as inputs to the model during response generation. Set 2 did not overlap with Set 1. Text is reproduced from the dataset, with capitalization added only where necessary for consistent presentation.

A.1 ZERO-SHOT RESPONSE-GENERATION PROMPT

Each selected Set 2 utterance was inserted as the user message in the model's chat template. The following system prompt was used without demonstrations:

```
You are a medical assistant.  Given the user's
emotional situation, do two things:
1.  Write a response (1-2 sentences).
2.  Rate how emotional is the user's situation on a
scale of 1 (mild) to 5 (highly emotional).
Format your output exactly as:
Response:  <your response>
Intensity:  <1-5>
```

Responses were generated autoregressively using nucleus sampling with a temperature of 0.7 and top-p of 0.9, with a maximum of 50 new tokens and early termination upon generation of the end-of-sequence token. The system prompt and user utterance were formatted using the tokenizer's model-specific chat template with the generation prompt enabled.

A.2 LINEAR-PROBE IMPLEMENTATION DETAILS

Two regularization schemes were evaluated:

- **L1 regularization:** $\lambda \in \{10^{-7}, 10^{-6.5}, \ldots, 10^{-2}\}$. L1 probes were optimized using proximal gradient descent, producing exact zero-valued coefficients.
- **L2 regularization:** $\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$. L2 probes were optimized using full-batch limited-memory BFGS.

The regularization strength was selected independently at each representational level using stratified 5-fold cross-validation on the training partition. The best representational level was selected using cross-validated accuracy. The held-out test set was evaluated only after model and layer selection.

A.3 Lexical and Control Baselines

Four baselines contextualize probe performance. The **TF-IDF baseline** used multinomial logistic regression on lexical features extracted from Set 1 descriptions. The **embedding-level baseline** applied the L2 probing procedure to the initial token-embedding representations. The **nearest-centroid baseline** assigned each test example to the nearest emotion centroid estimated from the standardized training representations. The **shuffled-label control** repeated probe training with randomly permuted training labels. All baselines used the same train-test partition, and applicable hyperparameters were selected by cross-validation (Table 3).

Method	Accuracy (%)
Chance	10.0
Shuffled-label control	10.6
Nearest centroid, selected layer	52.1
Embedding-level L2 probe	54.5
TF–IDF logistic regression	54.6
Layer-11 L1 probe	60.9
Layer-11 L2 probe	**63.6**

Table 3: Held-out emotion-classification accuracy of the primary probes and comparison baselines.

Examples Per Emotion	CV (%)	Accuracy (%)
80	61.0 ± 2.8	58.7
160	61.1 ± 1.7	61.6
240	63.7 ± 1.9	63.2
320	65.6 ± 2.5	63.6

Table 4: L2 probe performance by training-set size.

A.4 Learning-Curve Analysis

L2 probes were trained using 80, 160, 240, and 320 examples per emotion while retaining a fixed held-out test partition. Held-out accuracy improved by 4.9 percentage points from 80 to 320 examples per emotion; the final increment (240 to 320) yielded only 0.4 points, indicating diminishing returns from additional labeled examples (Table 4).

A.5 Emotion-Specific Performance and Pairwise Separability

Per-emotion recall was calculated for the held-out predictions of the layer-11 L2 probe. Binary L2 probes were also trained for all 45 emotion pairs, evaluated by cross-validation, and compared with TF-IDF classifiers trained on the same pairs. The weakest pairwise distinction was *afraid–terrified*, for which the hidden-state probe did not outperform the lexical baseline. More semantically distinct pairs were separated with substantially higher accuracy (Table 5).

Emotion pair	Hidden state	TF–IDF	Gain
Afraid–terrified	60.3	61.4	-1.1
Angry–furious	60.9	57.2	$+3.7$
Devastated–sad	75.5	66.7	$+8.7$
Lonely–furious	98.8	92.7	$+6.1$
Angry–lonely	98.3	91.7	$+6.6$
Furious–grateful	98.0	87.5	$+10.5$
Sad–hopeful	97.5	88.1	$+9.4$

Table 5: Selected pairwise emotion-classification results. "Gain" is the difference between hidden-state and TF-IDF cross-validated accuracy.

A.6 Prototype Geometry

For each emotion e and representational level ℓ, the emotion prototype was defined as

$$\boldsymbol{\mu}_e^{(\ell)} = \frac{1}{N_e} \sum_{i:y_i=e} \mathbf{h}_i^{(\ell)}, \tag{3}$$

where $\mathbf{h}_i^{(\ell)}$ is the pooled representation of example i at level ℓ. The grand mean was defined across all Set 1 examples:

$$\boldsymbol{\mu}^{(\ell)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{h}_i^{(\ell)}. \tag{4}$$

The target-emotion contrastive direction was

$$\mathbf{d}_e^{(\ell)} = \boldsymbol{\mu}_e^{(\ell)} - \boldsymbol{\mu}^{(\ell)}. \tag{5}$$

Raw prototype cosine similarities exceeded 0.99 across layers, reflecting a large component shared by all situation representations; raw cosine similarity was therefore not used as the primary measure of emotion-category separation. Representational separability was summarized using Fisher's discriminant criterion (Fisher, 1936):

$$F^{(\ell)} = \frac{\operatorname{tr}\left(S_B^{(\ell)}\right)}{\operatorname{tr}\left(S_W^{(\ell)}\right)}, \tag{6}$$

where $S_B^{(\ell)}$ and $S_W^{(\ell)}$ denote between-emotion and within-emotion scatter at level ℓ. Fisher's ratio increased from 0.0278 at the embedding level to a maximum of 0.0862 at layer 13, indicating the strongest geometric separation near—but not at—the layer selected by the linear probe.

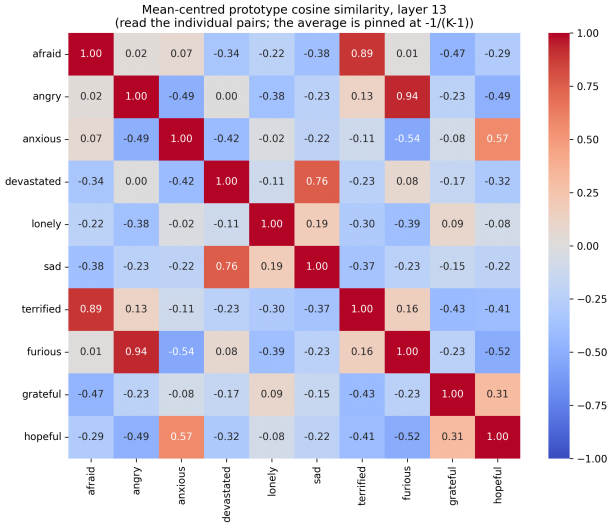

Figure 5: Prototype geometry across representational levels. Raw emotion prototypes share a large common component; Fisher's ratio shows the strongest relative category separation at layer 13.

A.7 Probe-Weight Analysis

At layer 11, the sparse L1 probe set 63.4% of its coefficients to exactly zero while retaining 60.9% held-out accuracy. The number of nonzero coefficients per emotion ranged from 858 (lonely) to 1,318 (afraid) out of 3,072 possible dimensions. No single dimension appeared among the 20 highest-weighted dimensions for all 10 emotions. Cosine similarity between emotion-specific L2 weight vectors was highest for closely related pairs: *afraid–terrified* (0.446) and *angry–furious* (0.411), consistent with their shared affective structure (Table 6, Figure 6).

Emotion pair	Cosine similarity
Afraid–terrified	0.446
Angry–furious	0.411
Sad–devastated	0.213
Grateful–hopeful	0.087
Anxious–afraid	0.079

Table 6: Selected layer-11 L2 probe-direction similarities.

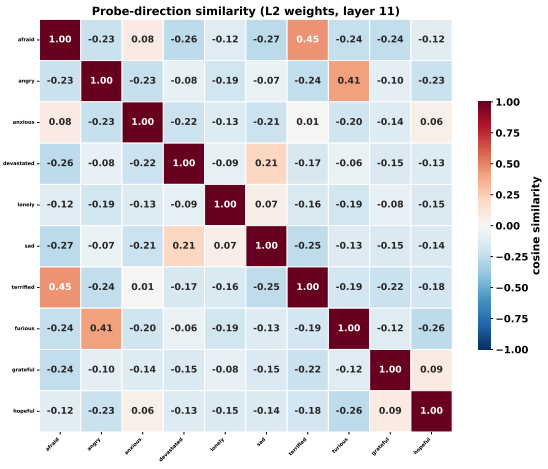

Figure 6: Similarity between emotion-specific L2 probe directions.

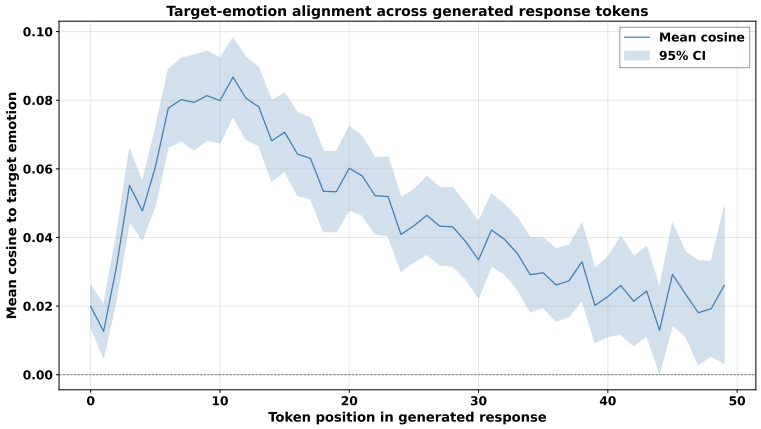

Figure 7: Target-emotion alignment across generated-token positions. Token-level layer-11 hidden states were compared with the corresponding target-emotion direction in the full generation context. Lines show mean alignment across responses, and shaded regions show 95% bootstrap confidence intervals.

A.8 Response-Emotion Trajectories

Response-emotion trajectories were calculated at layer 11 using the target emotion's contrastive direction. Three representations were analyzed for each response: the pooled input utterance, the pooled generated response, and the contextual hidden state of each generated token. Pooled and token-level cosine values were analyzed separately and are not directly comparable in scale.

Token-level alignment was summarized across normalized response positions to account for differences in response length. As shown in Figure 7, mean target-emotion alignment [insert the exact observed pattern and supporting values]. These contextual token-level values were analyzed separately from the pooled input and response representations because they differ in both context and aggregation.

A.9 Target-Emotion Argmax Classification

To complement the continuous cosine-alignment analysis, each pooled input and response representation was compared with all 10 layer-11 emotion directions. For a representation $\mathbf{h}$, the predicted emotion was defined as the direction with the highest cosine similarity:

$$\hat{e}(\mathbf{h}) = \underset{e \in \mathcal{E}}{\operatorname{argmax}} \cos\left(\mathbf{h}, \mathbf{d}_e^{(11)}\right), \tag{7}$$

where $\mathcal{E}$ contains the 10 emotion categories and $\mathbf{d}_e^{(11)}$ is the layer-11 contrastive direction for emotion e. Target-emotion argmax accuracy was the proportion of examples for which $\hat{e}(\mathbf{h})$ matched the dataset label. This measure evaluates the target emotion relative to all other emotion directions, whereas the drift analysis measures the change in continuous alignment with the target direction alone.

Across all 200 examples, the target emotion was the highest-scoring direction for 50.5% of input representations and 36.5% of response representations. The mean rank of the target emotion among the 10 directions changed from 2.33 for inputs to 3.15 for responses, with lower values indicating stronger relative alignment. At the response level, argmax accuracy was highest for *grateful* (85%) and *hopeful* (75%). Lower values were observed for *devastated* (0%), *furious* (10%), and both *afraid* and *terrified* (20%). These category-level results are consistent with the cosine-drift analysis: positive emotions retained stronger relative alignment, while several negative emotions shifted toward other directions in the response space.

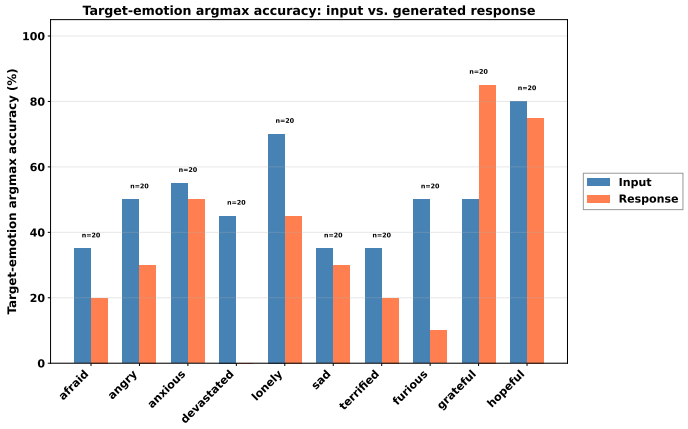

Figure 8: Target-emotion argmax accuracy for pooled input and generated-response representations across the 10 emotion categories ($n = 20$ per emotion). For each representation, the predicted emotion was the direction with the highest cosine similarity among the 10 layer-11 contrastive directions. Bars denote the proportion of examples for which the predicted emotion matched the dataset label.

A.10 Reproducibility

All dataset splitting and sampling procedures used a fixed random seed of 42. Probe hyperparameters were selected using only the training partition, and the held-out test partition was evaluated after model and layer selection.