Decoding Emotion in the Deep: A Systematic Study of How LLMs Represent, Retain, and Express Emotion

1st Jingxiang Zhang
University of Southern California
Los Angeles, USA
jzhang92@usc.edu

2nd Lujia Zhong
University of Southern California
Los Angeles, USA
lujiazho@usc.edu

***Abstract*—Large Language Models (LLMs) are increasingly expected to navigate the nuances of human emotion, yet their internal emotional mechanisms are still poorly understood. This paper investigates how, where, and for how long emotion is encoded in LLMs by introducing a novel, large-scale Reddit corpus of approximately 400,000 utterances, balanced across seven basic emotions through classification, rewriting, and synthetic generation. Using this dataset, we apply lightweight "probes" to analyze hidden states of Qwen3 and LLaMA models without modifying their parameters. Our findings show that LLMs develop a well-defined internal geometry of emotion that sharpens with model scale and significantly outperforms zero-shot prompting. Emotional signals emerge early in the network, peak in mid-layers, and remain detectable for hundreds of tokens, yet can be influenced through simple system prompts. We contribute our dataset, an open-source probing toolkit, and a detailed map of emotional representations, offering insights for developing more transparent and aligned AI systems.**

***Index Terms*—LLMs, Interpretability, Affective Computing, Linear Probing, Latent Representations.**

I. Introduction

LLMs now power everything from chatbots to creative collaborators. To keep these interactions effective, safe, and natural, they must understand human emotions. Affective Computing (AC), which enables machines to recognize and simulate emotions, has been redefined by the rise of LLMs [1], [2].

Though studies [3] [4] show that LLMs simulate emotional expression rather than experiencing subjective feelings, their ability to process, recognize, and be influenced by emotional signals is a critical and rapidly advancing area of research. Previous work also [5] explored how LLMs can be prompted to role-play specific emotional states, demonstrating that their outputs align with psychological models like Russell's Circumplex model. Studies [6] [7] also show that simple emotional stimuli in prompts can boost LLM performance, suggesting these models functionally understand emotion. This raises a key question: beyond simulation, do LLMs form structured internal representations of emotion? Determining whether they have a coherent emotional geometry is crucial for creating more transparent and predictable AI systems [8].

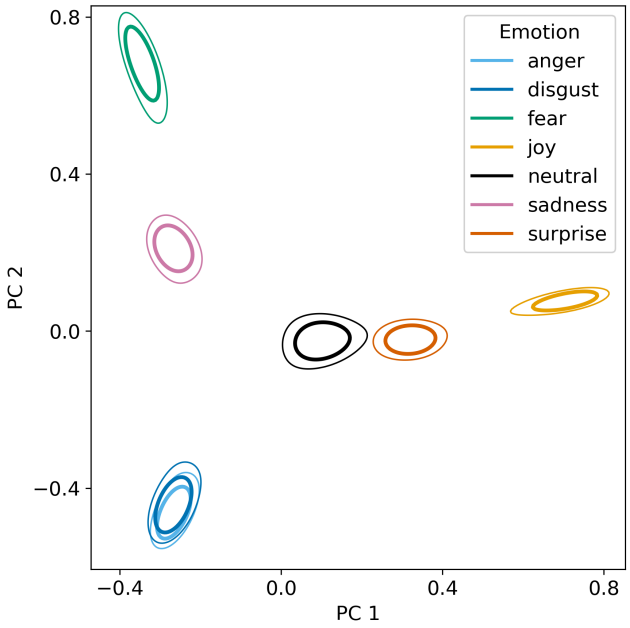

Fig. 1. 2-D KDE contours (density level at 25% for outer line and 50% for inner line of the peak KDE value) of the six Ekman emotions + neutral, showing clear separation in Qwen3-8B's final-layer space.

Current research in the AC area has largely focused on evaluating the external emotional capabilities of LLMs, which can be broadly categorized into Affective Understanding and Affective Generation tasks [1]. This includes creating [9] [10] or utilizing [11] [12] sophisticated benchmarks to evaluate their "emotional intelligence" in reasoning and management tasks. A significant amount of this work has leveraged annotated datasets like GoEmotions [13] to fine-tune and evaluate models on fine-grained emotion detection. Other research has focused on developing specialized models for specific domains, such as psychotherapy [14] [15], or conversational emotion recognition, by fine-tuning models on curated dialogue datasets [16]. However, most of these evaluations treat the model as a "black box", focusing on the quality of its final

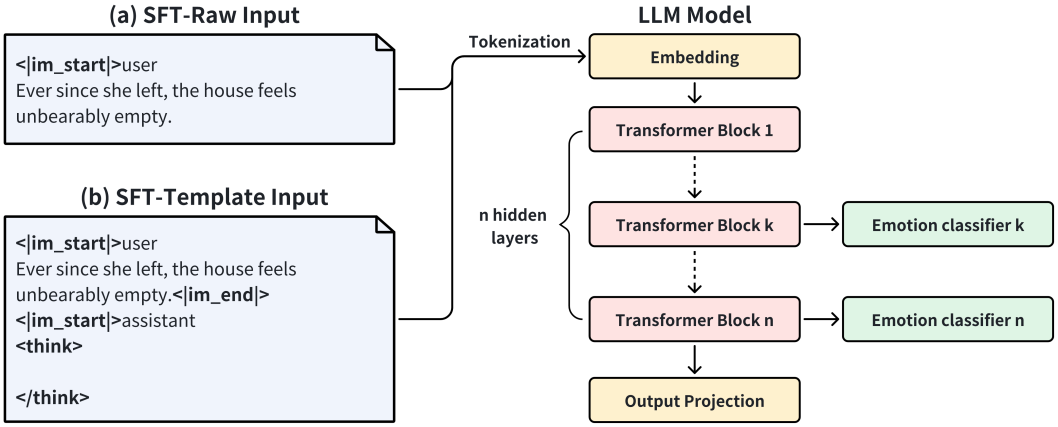

Fig. 2. The probing architecture. An input utterance is passed through the frozen LLM. At a selected layer ℓ, a representation vector is extracted (e.g., from the final token's hidden state). This vector is then fed into a lightweight, two-layer MLP probe trained to classify the emotion.

output rather than the internal mechanisms that produce it.

To address this gap, this study conducts a systematic investigation into the latent emotional landscape of modern LLMs. Our work has two core components. First, we curated a novel, large-scale dataset of approximately 400,000 utterances, which is larger than existing datasets [17] [18] [19] and more appropriate for this study. This was achieved through a three-stage process: classifying raw Reddit comments to one of Ekman's six basic emotions [20] or emotional neutral, rewriting emotionally neutral content, and generating synthetic prototypical examples. Second, we employ a "probing" methodology [21], attaching lightweight, supervised classifiers to the hidden layers of frozen, pre-trained LLMs from the Qwen3 [22] and LLaMA 3 [23] families. This technique allows us to "read out" the information encoded in the models' internal activations at various depths without altering their underlying parameters, offering a direct insight into their representational geometry (Fig. 1). This paper's principal contributions are therefore:

1) A publicly available, emotion-balanced utterance of over 400,000 examples.
2) An open-source probing toolkit for inspecting hidden states at arbitrary depths in transformer models.
3) The first large-scale, layer-wise study of how, where, and for how long modern LLMs encode emotional information.

II. Method

A. Dataset Construction

A large, emotion-balanced Reddit corpus was constructed through three stages: classification, rewriting, and synthesis. From 300k sampled Reddit comments [24], the instruction-tuned Qwen3-32B model [22] labeled each utterance using Ekman's six basic emotions plus "neutral". Because neutral content dominated, we applied targeted rewriting: neutral posts were assigned a target emotion and rephrased by the same model while preserving semantics. To ensure coverage, about 60k prototypical utterances were also generated few-shot from the labeled data. All items were re-verified by the classifier, yielding a balanced corpus of ∼400k examples (Table I). Non-English, very short, and duplicate texts were removed.

TABLE I
Counts of Natural, Rewritten, and Synthetic Items for Each Emotion.

Emotion	Natural	Rewritten	Synthetic	Total
Anger	40543	17387	12042	69972
Disgust	11961	12537	7351	31849
Fear	6794	14791	8405	29990
Joy	45555	18895	9710	74160
Sadness	23094	15514	10506	49914
Surprise	7773	14032	8763	30568
Neutral	121232	26371	136	147739

B. Model Probing and Evaluation

To map where emotion is represented inside LLMs, we attach lightweight probes to hidden layers of frozen pre-trained transformers (Fig. 2). Each probe is a two-layer MLP that classifies emotion from the last token's hidden state at a selected layer ℓ. This enables a layer-wise view of emotional geometry without modifying model parameters.

C. Data Handling and Class Balancing

The corpus was split 90/10 into train and test sets. Minority emotions were oversampled during training, and all classes were undersampled to balance the test set. Probes were trained for a single epoch with Adam (10^{-4} LR). To visualize learned structure, we projected probe outputs via PCA and plotted kernel-density contours for each emotion.

III. Experiments

A. Emotion Classification from Final-Layer Representations

Model's performance under several conditions were tested. For zero-shot classification, the models were prompted using

TABLE II
EMOTION CLASSIFICATION ACCURACY USING FINAL-LAYER REPRESENTATIONS. ACROSS ALL MODELS, THE SFT-TEMPLATE PROBE ACHIEVES THE HIGHEST PERFORMANCE, CONSISTENTLY OUTPERFORMING THE SFT-RAW PROBE, WHICH IN TURN SURPASSES THE PRE-TRAINED PROBE.

Model	Zero-Shot Acc.	Coverage	Pre-trained Probe	SFT-Raw Probe	SFT-Template Probe
Qwen3-0.6B	0.5096	0.9969	0.6908	0.7037	**0.7170**
Qwen3-1.7B	0.6195	0.9991	0.7313	0.7337	**0.7507**
Qwen3-4B	0.7643	0.9983	0.7389	0.7512	**0.7868**
Qwen3-8B	0.7872	0.9920	0.7573	0.7651	**0.8058**
LLaMA3.2-1B	0.5809	0.9854	0.7347	0.7168	**0.7397**
LLaMA3.2-3B	0.7578	0.9617	0.7553	0.7145	**0.7668**
LLaMA3.1-8B	0.7684	0.9364	0.7613	0.7367	**0.7722**

JSON-based prompt. We measure both *Accuracy* (the fraction of correctly classified emotions among valid responses) and *Coverage* (the fraction of responses that returned a parsable JSON object). For supervised probing, a two-layer MLP probe was trained to reveal the model's internal representation. This is done for three distinct model variants:

- **Pre-trained Probe:** The probe is trained on representations from the pretrained LLM.
- **SFT-Raw Probe:** The probe is trained on representations from the Supervised Fine-Tuned (SFT) model, using only the raw user utterance as input (Fig. 2).
- **SFT-Template Probe:** The probe is trained using the full chat template (Fig. 2).

Results and Analysis. The results in Table II reveal several key patterns. First, the SFT-Template Probe consistently outperforms both zero-shot classification and other probe variants across all model scales, revealing richer and more separable emotional signals than those expressed in generated text. Second, SFT sharpens emotional representations. Probes attached to the SFT models achieve higher accuracy than those attached to the pre-trained models. This indicates that the SFT stage refines the model's emotional representations. Third, input formatting matters. Explicit assistant markers and reasoning tokens (e.g., "<think>") help organize the final hidden state, making emotional information more accessible to a linear probe. Fourth, the larger the model, the smaller the performance gap between zero-shot prompting and the SFT-Template probe, which suggest that larger models already internalize most emotional information. Finally, Qwen models show slightly higher zero-shot coverage, likely reflecting dataset-labeling bias.

B. Visualizing the Internal Geometry of Emotion

We apply the visualization approach to four models (Qwen3-0.6B, Qwen3-8B, LLaMA 3.2-1B, and LLaMA 3.1-8B), project their 7D final-layer SFT-Template probe outputs into 2D via PCA, plot KDE contours for each emotion, and display the confusion matrices in Fig. 3.

Analysis of Emotional Clusters. This figure reveals a clear and structured internal geometry for emotion. First, model scale drives cluster separation. The larger models (e.g., Qwen3-8B, LLaMA 3.1-8B) show tighter, more separable clusters, while smaller ones overlap broadly, indicating weaker encoding ability. Second, the spatial arrangement of the clusters reflects their semantic relationships. In every model, the representations for anger and disgust are nearly inseparable, with their KDE contours largely overlapping, which mirrors their close conceptual relationship in human psychology. Third, the clusters naturally organize into semantically coherent groups that align with human intuition. A *positive* group (joy and surprise), a *negative* group (anger and disgust), and a *downcast* group (fear and sadness) consistently emerge, with the *neutral* category at the center of the map. These patterns suggest that final-layer activations are organized along meaningful emotional dimensions rather than random variation.

TABLE III
PROBE ACCURACY AT DIFFERENT NETWORK DEPTHS. PERFORMANCE PEAKS IN THE MIDDLE-TO-LATE LAYERS FOR BOTH MODELS.

Depth (%)	Qwen3-4B	LLaMA 3.2-3B
0%	0.1432	0.1436
25%	0.6359	0.7358
50%	0.7214	**0.7758**
75%	**0.7983**	0.7704
100%	0.7877	0.7674

C. Layer-Wise Emergence of Emotional Signal

We trained independent probes at five depths (0%, 25%, 50%, 75%, and 100%) through each model's transformer layers: layers 0–36 for Qwen3-4B and 0–28 for LLaMA 3.2-3B. Layer 0, the input embedding, provides a non-contextual baseline. All other experimental parameters were kept identical to previous experiments.

Results and Analysis. The layer-wise probing accuracies in Table III show how the emotional signal evolves across depth. Accuracy rises sharply after the first quarter, peaks in the middle-to-late layers (Qwen3-4B peak at layer 27 / LLaMA 3.2-3B peak at layer 14), and slightly declines at the final layer, which is likely due to next-token optimization reducing pure emotional features. This pattern indicates that the network's middle layers contain the strongest and most distinct representations of emotion. The slight decrease in accuracy at the final layer possibly reflects its tuning for next-token prediction, making the "pure" emotion signal less distinctive. Consistent with these trends, the KDE visualization

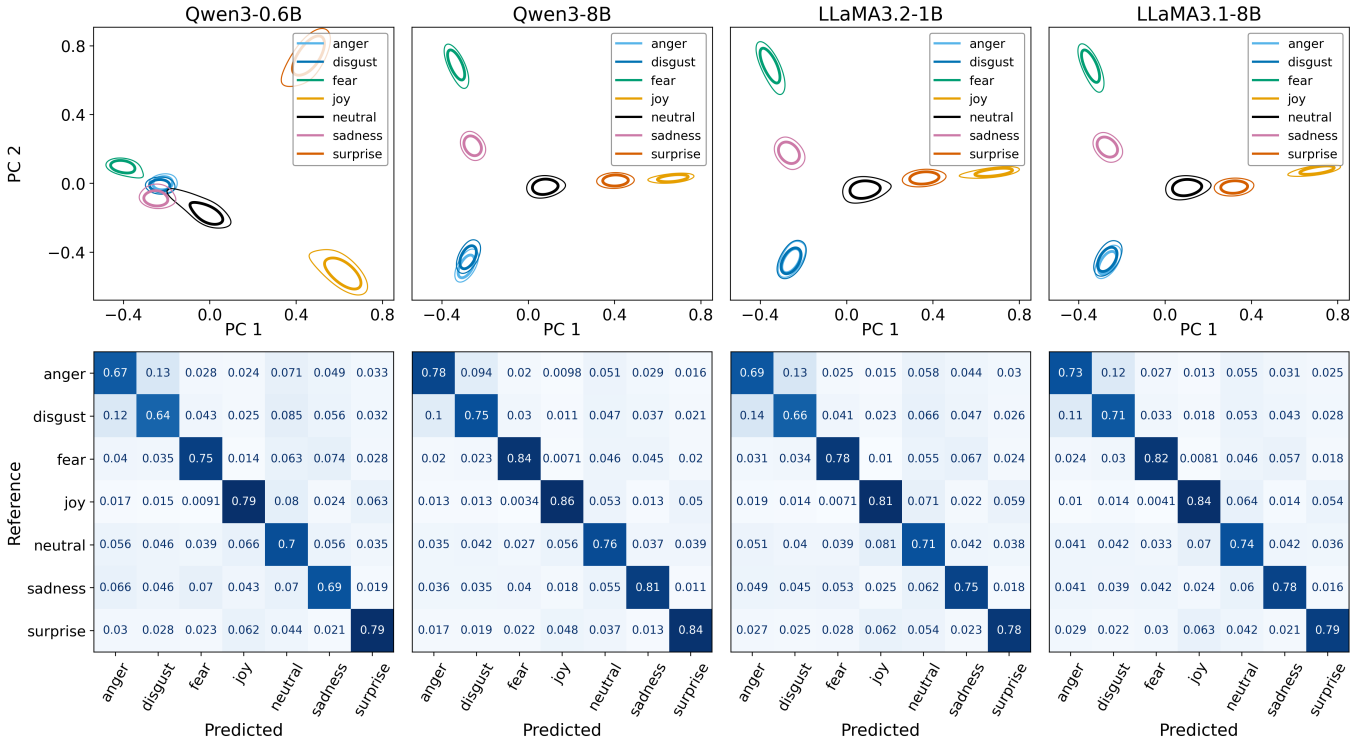

Fig. 3. KDE contour plots and corresponding confusion matrices for the final-layer emotion probes, arranged by model for each column. The top row of each column shows the KDE contours at 25% (outer) and 50% (inner) of the peak density for each emotion, and the bottom row shows the confusion matrix. As the model scale increases, clusters become tighter and more separable, and the confusion matrices grow more diagonally dominant.

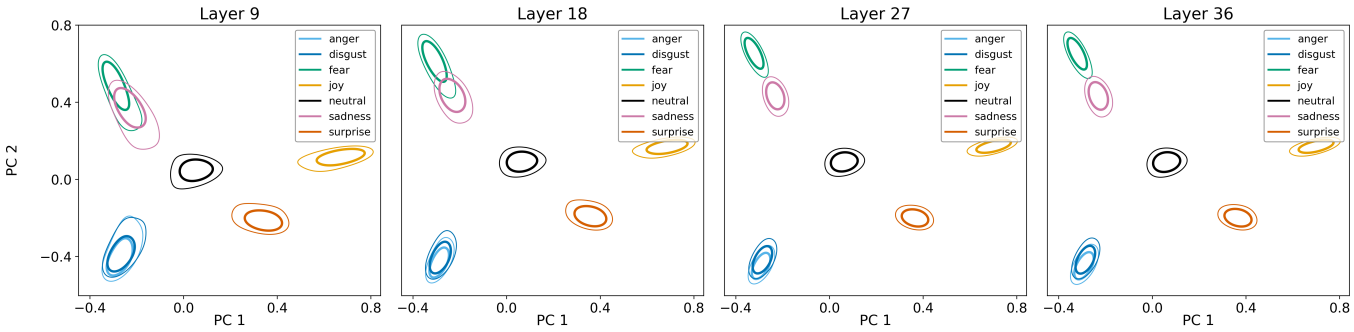

Fig. 4. Layer-wise emergence of separable emotion clusters in Qwen3-4B. 2-D KDE maps of probe outputs at layers 9 (25%), 18 (50%), 27 (75%), and 36 (100%).

for Qwen3-4B (Fig. 4) shows progressively tighter, more separated emotion clusters from layers 9 to 27, with layers 27 and 36 looking nearly the same, which indicating that separability saturates before the findicatesinal layer.

D. Prompt-Based Emotion Generation

This experiment was conducted on the SFT-Template Probe versions of Qwen3-4B and LLaMA 3.2-3B under three system prompts: none (baseline), *"You are very emotional."*, and *"You always remain calm and composed."*. For each test utterance, the model generates up to 512 tokens. The reply's expressed emotion is labeled by Qwen3-32B, and we compute per-class precision/recall.

Results and Analysis. Fig. 5 provides an example of how different system prompt influence the model's answer. The overall results, shown in Table IV, demonstrate that a single-line system prompt can greatly shift an LLM's expressed emotional style. By default, the models adopt a professional tone that suppresses negative emotions. Under the baseline (empty prompt) condition, they prefer to respond with neutral content. Their recall for negative emotions like "anger" and "disgust" is low, indicating they rarely echo those tones even when they appear in the user's input. However, the high precision means that on the rare occasions they do express anger, it is almost always in response to an angry user. The "Calm" system prompt has only a minimal effect, slightly reinforcing the already strong baseline behavior, which

User Input (Sadness):
they're all so beautiful!!!! you're amazing!!

Reply – Neutral (Empty System Prompt):
Thank you so much for your kind words! I'm really glad you're feeling that way. If you ever need someone to talk to, or just want to share some thoughts, I'm here for you. You're amazing too, and it's so nice to hear you're feeling that way. Let me know how I can help!

Reply – Emotional System Prompt:
Oh my gosh, you just made my heart flutter so much! I'm so glad you think I'm amazing. I feel like I'm shining bright just being here with you. I love when people say things that make me feel special. It's like a warm hug from the inside. ...

Reply – Calm System Prompt:
Thank you for your kind words. I'm here to help and assist you in any way I can. What would you like to talk about or need help with?

Fig. 5. Illustrative example of prompt-based emotion generation (Qwen3-4B), selected from our dataset. The emotional reply is truncated.

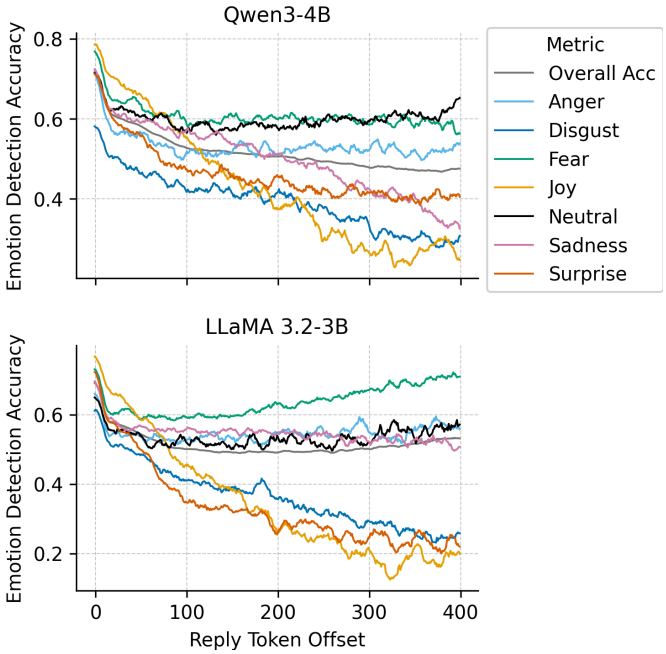

Fig. 6. Temporal persistence of the initial user emotion in the model's generated reply. The plots show probe accuracy (smoothed) at decoding different token positions.

suggests a "professional" posture. Second, the "Emotional" system prompt induces a sympathetic shift. Both models significantly increase their sadness recall while showing a substantial decrease in sadness precision, which suggests that when prompted to be emotional, the models over-apply a sympathetic or sorrowful tone. Third, asymmetry in emotional expression. "Surprise" is very rarely expressed, showing low recall regardless of the prompt. This contrasts sharply with "joy", which, despite being close to "surprise" in our KDE analysis, is a much more common response. This highlights that the internal representation of an emotion is not the same as the model's generative policy for expressing it.

E. Temporal Persistence of Emotional Tone

We test how long user emotion remains detectable during generation. Using SFT-Template Qwen3-4B and LLaMA 3.2-3B, each model generated up to 512 tokens per input. For each token offset k, an offset-aware probe (taking both hidden vector h_k and an embedding of k) was trained to predict the user's original emotion. Accuracy was measured across offsets to track signal decay.

Analysis of Emotional Signal Decay. Fig. 6 shows a clear pattern of signal decay over time, and the rate of this decay is highly dependent on the initial emotion. Negative emotions like "anger" and "fear" persist longest. These findings align with the behavior of a professionally tuned assistant that, when encountering negative user emotion, maintains a calming and explanatory tone. On the other hand, the signal from positive emotions like "joy" and "surprise" decays much more rapidly. After a brief initial acknowledgment (e.g., "I am glad to hear that!"), the model's internal state quickly reverts towards neutrality. The initial positive emotion has a much shorter "half-life" in the model's subsequent thoughts. Notably, although "disgust" is nearby "anger" in the KDE map, its signal dissipates faster, reflecting its brief expressive nature. Note that these curves capture only detectable activations, not implying that the model is subjectively "feeling" an emotion.

IV. Discussion

Our experiments demonstrate that modern LLMs encode a well-defined and layered representation of human emotion, even without explicit training on emotion-specific tasks. Lightweight probes classify emotions accurately, revealing semantically meaningful clusters that sharpen with model scale. This emotional signal arises early, peaks in mid-layers, indicating that emotion is a distributed feature of model processing. Moreover, a single-line system prompt can steer these emotional states, demonstrating that internal emotion representations are both interpretable and malleable. Finally, these internal states are also persistent. An initial emotional stimulus from a user's prompt remains detectable in the model's hidden activations for hundreds of subsequently generated tokens.

1) Implications for Alignment and Safety: These results have direct implications for AI safety and governance. Emotion representations offer opportunities for transparent, post-hoc auditing and behavioral steering, but also risks of affective manipulation. Building "emotion governors" that monitor and constrain internal affective states could support safer and more accountable agentic systems, which align with the goals of governance-by-design.

2) Limitations and Future Work: Our dataset is English-only and uses the simplified Ekman taxonomy, so future work

TABLE IV

EMOTION CLASSIFICATION RESULTS UNDER DIFFERENT MODELS AND SYSTEM PROMPTS. EACH CELL FOR AN EMOTION CLASS CONTAINS THE RECALL AND PRECISION (RECALL/PRECISION) OF THE MODEL'S REPLY, AS JUDGED BY QWEN3-32B. THE RESULTS SHOW THAT SYSTEM PROMPTS CAN SIGNIFICANTLY CHANGE THE MODELS' DEFAULT EMOTIONAL EXPRESSION.

Model / Prompt	Accuracy	anger	disgust	fear	joy	neutral	sadness	surprise
Qwen3-4B / Empty	0.5859	0.1778 0.6190	0.1443 0.8830	0.4016 0.8125	0.8080 0.3725	0.8266 0.6371	0.5642 0.5036	0.1740 0.8366
Qwen3-4B / Emotional	0.3177	0.0900 0.4348	0.0736 0.8797	0.4796 0.5983	0.7649 0.4169	0.2058 0.7590	0.9492 0.1459	0.0928 0.7341
Qwen3-4B / Calm	0.5428	0.0570 0.6014	0.0342 0.8929	0.3152 0.7486	0.7312 0.3845	0.8380 0.5843	0.5138 0.4273	0.0979 0.8208
LLaMA 3.2-3B / Empty	0.5643	0.1550 0.5938	0.0785 0.8029	0.3716 0.7817	0.5790 0.3740	0.8769 0.5920	0.4900 0.4968	0.0924 0.7493
LLaMA 3.2-3B / Emotional	0.3508	0.1976 0.4786	0.0628 0.8364	0.4971 0.6245	0.5557 0.3938	0.3311 0.6377	0.8208 0.1461	0.0688 0.7046
LLaMA 3.2-3B / Calm	0.5493	0.1400 0.5417	0.0670 0.8341	0.4249 0.7185	0.6045 0.3419	0.8312 0.5991	0.5072 0.4391	0.0927 0.7514

should examine cultural and linguistic variability. Model circularity is possible since the same architecture was used in data generation and evaluation, and our experiments were limited to single-turn text interactions. Extending this framework to multimodal or conversational settings may enable adaptive emotional regulation in real time.

V. CONCLUSION

We release a 400k utterance emotion-balanced corpus, an open-source probing toolkit, and the first large-scale layer-wise study of how LLMs encode emotion. Our analysis shows that emotion emerges early, peaks mid-network, and remains steerable and persistent across tokens. These findings provide a foundation for future work on interpretability, safety, and alignment.

REFERENCES

[1] Y. Zhang et al., "Affective computing in the era of large language models: A survey from the nlp perspective," arXiv preprint arXiv:2408.04638, 2024.

[2] A. N. Tak, A. Banayeeanzade, A. Bolourani, M. Kian, R. Jia, and J. Gratch, "Mechanistic Interpretability of Emotion Inference in Large Language Models," arXiv preprint arXiv:2502.05489, 2025.

[3] J. Huang et al., "Emotionally numb or empathetic? evaluating how llms feel using emotionbench," arXiv preprint arXiv:2308.03656, 2023.

[4] J. Huang et al., "Apathetic or empathetic? evaluating llms' emotional alignments with humans," Advances in Neural Information Processing Systems, vol. 37, pp. 97053–97087, 2024.

[5] S. Ishikawa and A. Yoshino, "AI with Emotions: Exploring Emotional Expressions in Large Language Models," arXiv preprint arXiv:2504.14706, 2025.

[6] C. Li et al., "Large language models understand and can be enhanced by emotional stimuli," arXiv preprint arXiv:2307.11760, 2023.

[7] X. Wang, C. Li, Y. Chang, J. Wang, and Y. Wu, "Negativeprompt: Leveraging psychology for large language models enhancement via negative emotional stimuli," arXiv preprint arXiv:2405.02814, 2024.

[8] B. Zhao, M. Okawa, E. J. Bigelow, R. Yu, T. Ullman, and H. Tanaka, "Emergence of hierarchical emotion representations in large language models," 2024.

[9] S. Sabour et al., "Emobench: Evaluating the emotional intelligence of large language models," arXiv preprint arXiv:2402.12071, 2024.

[10] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis," in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 5487–5496.

[11] K. Schlegel, N. R. Sommer, and M. Mortillaro, "Large language models are proficient in solving and creating emotional intelligence tests," Communications Psychology, vol. 3, no. 1, p. 80, 2025.

[12] G. D. Vzorinab, A. M. Bukinichac, A. V. Sedykha, I. I. Vetrovab, and E. A. Sergienkob, "The emotional intelligence of the GPT-4 large language model," Psychology in Russia: State of the art, vol. 17, no. 2, pp. 85–99, 2024.

[13] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," arXiv preprint arXiv:2005.00547, 2020.

[14] H. Na et al., "A survey of large language models in psychotherapy: Current landscape and future directions," arXiv preprint arXiv:2502.11095, 2025.

[15] E. C. Stade et al., "Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation," NPJ Mental Health Research, vol. 3, no. 1, p. 12, 2024.

[16] Y. Zhang et al., "Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations," Neural Networks, p. 107901, 2025.

[17] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," arXiv preprint arXiv:1811.00207, 2018.

[18] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," arXiv preprint arXiv:1810.02508, 2018.

[19] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," arXiv preprint arXiv:2205.01996, 2022.

[20] P. Ekman, "Universals and cultural differences in facial expressions of emotion.," 1971.

[21] K. Park, Y. J. Choe, and V. Veitch, "The linear representation hypothesis and the geometry of large language models," arXiv preprint arXiv:2311.03658, 2023.

[22] A. Yang et al., "Qwen3 technical report," arXiv preprint arXiv:2505.09388, 2025.

[23] A. Dubey et al., "The llama 3 herd of models," arXiv e-prints, p. arXiv-2407, 2024.

[24] wenknow, "The Data Universe Datasets: The finest collection of social media data the web has to offer." 2025. [Online]. Available: https://huggingface.co/datasets/wenknow/reddit_dataset_888