Simulating Emotional Intelligence in LLMs through Behavioral Conditioning and Analogical Retrieval

G. Sai Linisha Reddy **Mounil Hiren Kankhara** **Mridul Maheshwari** **Swayam Bansal**
Rishit Kapoor **Himesh Reddy M** **Bagesh Kumar**
Manipal University Jaipur, India
g.229302590@muj.manipal.edu, mounil.23fe10ite00021@muj.manipal.edu,
mridul.23fe10ite00041@muj.manipal.edu, swayam.23fe10ite00308@muj.manipal.edu,
rishit.23fe10ite00026@muj.manipal.edu, himesh.229303296@muj.manipal.edu,
pse2016001@iiita.ac.in

Abstract

Human emotional expression emerges from a complex interplay of verbal, para-verbal, and non-verbal cues. This paper presents a dual-path framework for emotionally grounded text generation in large language models by integrating behavioral metadata with analogical retrieval. We introduce the MECC (Multimodal Emotionally Conditioned Corpus), a dataset of 1,764 question-answer pairs collected via structured interviews and annotated across 15 emotion categories with tone, response time, and body language. A LLaMA-3.1–8B–Instruct model is fine-tuned on MECC using behavior-encoded prompts, and inference is supported by a metadata-filtered Retrieval-Augmented Generation (RAG) pipeline. Detailed emotion-level analysis reveals trade-offs between emotional fidelity and semantic diversity, emphasizing the need for nuanced evaluation. This study contributes a richly annotated multimodal emotion corpus, a metadata-driven RAG architecture, a well-structured framework for building emotionally aware language models.Our code is available at https://github.com/MetaResearcher/Framework

1 Introduction

"*Emotion is not opposed to reason; it is its foundation.*" — Antonio Damasio, Descartes' Error

Human emotional reasoning is seldom simply associative or reactive. Rather, it is inherently analogical, shaped by past experiences, moderated by the present context, and shaped by memory, rich in metaphors for how they make decisions now. In emotionally charged circumstances, people often reflect analogically: "This is like how I felt when..." Indeed, this notion of analogical reasoning accounts for foundational cognitive models of analogy (Gentner, 1983)(Gärdenfors, 2000)(Holyoak and Thagard, 1995) that have been used to articulate concepts of affective computing and socially intelligent systems (Picard, 1997)(Hoegen et al., 2019)

Nonetheless, even with tremendous advancements in artificial intelligence and current generative models, the state-of-the-art is underwhelmed in its ability to replicate this level of nuance with emotional reasoning. Given that most Large Language Models (LLMs) leverage only surface-level conditioning, using emotion tokens (Zhou et al., 2018) sentiment prompts, or affective keywords (Rashkin et al., 2019), at the surface without translating the embodied, context-sensitive nature of a human emotional experience. Further, as with many machines initially designed for logic or limited predictability, LLMs miss integral non-verbal elements involved with human emotional messaging: tone, timing, gesture, etc. (Mehrabian, 1971)(Wang et al., 2004).

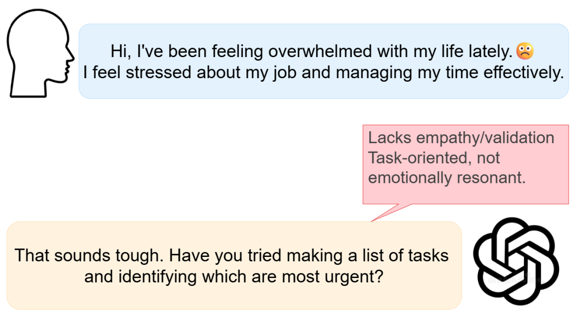

Figure 1: Figure of ChatGPT-4o response where it fails to generate an emotionally aware response

The degree of disconnection noted above is important in human-centered areas of work like healthcare, education, and therapy, in which emotionally coherent and capturing the unique personal connection is important. Recent applications using generative systems to enhance empathy in clinical documentation (Nag et al., 2023) and encourage pro-social engagement in environmental psychology (Lim et al., 2024), demonstrate that more complex cognitive-emotional capacities, such as

Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II), pages 81–91

introspection, empathy, and analogy, are lagging (Ortega-Ochoa et al., 2024)(Varma et al., 2024).

Building on these research gaps, we present the **MECC** (Multi-modal Emotionally Conditioned Corpus), a behaviorally enriched data set constructed from a 60-item psychological interview protocol. The data set includes 1,764 pairs (question, response) from 31 participants, annotated across 15 emotional categories and accompanied by behavioral metadata such as vocal tone, response time, and body language. By incorporating both verbal and paraverbal dimensions, MECC enables a more cognitively grounded investigation of affective states.

Using MECC, we fine-tuned the LLaMA-3.1–8B-Instruct model on behavior-encoded prompts and integrated a metadata-filtered Retrieval Augmented Generation (RAG) pipeline. This dual-path architecture conditions the generation on both situational context and analogically retrieved affective exemplars. The combined system goes beyond surface-level fluency by generating emotionally aligned responses that reflect behavioral and situational grounding, supporting more context-sensitive and affect-aware language generation.

2 Related Work

Recent developments in large language models (LLMs) have had a great impact on emotion understanding tasks generally, and multimodal emotion understanding tasks specifically. (Luo et al., 2024) experimented with several LLMs for a specific variant of the Multimodal Emotion Cause Pair Extraction with Emotion Category (MECPE-Cat) task. After noting that ChatGLM had the highest performance, they reported a weighted average F1 score of 34.71 percent using just two training epochs. Prompt engineering was an important aspect of this team's system, and they got their emotion-labeled data from the ECF dataset, which contained a total of 13,619 utterances, in an overwhelmingly constructive way with the data through the use of prompt engineering.

(Wang et al., 2024) expanded on Luo et al. 's work by running team submissions for two subtasks: text-based emotion-cause pair extraction (TECPE) and multimodal emotion-cause pair extraction (MECPE). For the task of TECPE, the highest F1 score reported was 0.3223 with a combination of LLaMA2 and SpanBERT. For the subtask of MECPE, the highest F1 score reported was 0.3774 with an ensemble of LLaMA2, RoBERTa, and LLaMA, showing that multimodal models tend to outshine unimodal models.

Nevertheless, standard unimodal methods may not offer the complexity necessary to successfully address authentic emotional expression. (Cheng et al., 2024) saw the importance of this and created the MERR dataset, which included 28,618 coarse-grained and 4,487 fine-grained annotated examples. Their Emotion-LLaMA model, through instruction tuning, achieved better performance than existing multimodal LLMs (MLLMs) across many benchmarks. They achieved the highest scores in Clue Overlap (7.83) and Label Overlap (6.25) on the EMER dataset, and an F1 score of .9036 on the MER2023-SEMI challenge, along with good zero-shot results on the DFEW dataset.

Another study of emotional intelligence in LLMs was conducted by (Chen et al., 2024) with EmotionQueen, a standardized benchmark containing 10,000 statements generated by GPT-4 across five life scripts. They examined 11 major LLMs on four tasks: Event Recognition, Mixed Event Recognition, Implicit Emotion Recognition, and Intention Recognition. The study found that LLaMA-70B obtained the highest average score (93.4), while Claude2 was ranked first in Key Event Recognition. While pre-trained attention-based LLMs such as GPT-4 can achieve high accuracy on certain emotion recognition tasks, they often fail to translate this accuracy into responses that demonstrate compassionate or emotionally supportive reasoning. That is, the models may correctly identify emotional cues but lack the ability to respond in a way that reflects genuine empathy or care.

In addition to these studies on recognition, (Varma et al., 2024) introduced an emotionally adaptive AI pipeline to generate personalized emotional responses. They compared a fine-tuned LLaMA-3 8B modulated by LoRA with an RAG system. The fine-tuned model trained using interview data containing six universal emotions outperformed the RAG system; all of their correlation indices were greater than 0.950 (i.e., 0.850 for Anger) with a lower MSE score overall, such as 0.0452. Their research supports the notion that RAG systems struggle with emotional nuance. This motivates our dual-path design that includes analogical retrieval and behavioral conditioning.

Overall, the recent advances in the literature mark an important shift towards multimodal LLMs

No.	Paper / Authors	Model/Method	Contribution
1	Luo et al. (2024)	ChatGLM + Prompt Engineering	Achieved 34.71% F1 on MECPE-Cat using ECF dataset; demonstrated effective prompt design.
2	Wang et al. (2024)	LLaMA2 + SpanBERT; Ensemble (LLaMA2, RoBERTa, LLaMA)	Reported highest F1 of 0.3223 (TECPE) and 0.3774 (MECPE); showed multimodal models outperform unimodal.
3	Cheng et al. (2024)	Emotion-LLaMA (Instruction Tuning)	Introduced MERR dataset; achieved state-of-the-art on multiple emotion benchmarks.
4	Chen et al. (2024)	11 LLMs (incl. LLaMA-70B, Claude2, GPT-4)	Proposed EmotionQueen benchmark; analyzed LLMs on emotional intelligence tasks.
5	Varma et al. (2024)	Fine-tuned LLaMA-3 8B (LoRA) vs. RAG	Developed emotionally adaptive pipeline; fine-tuned model outperformed RAG in nuance and accuracy.

Table 1: Summary of recent literature on LLMs for emotion understanding.

and instruction tuning, aiming to capture nuanced emotional understanding. However, there remains a notable gap in models that can accurately identify emotion cause relationships while simultaneously adapting empathetically to behavioral context and modality. This study addresses that gap through a behavior-conditioned framework grounded in analogical retrieval and affective reasoning.

3 MECC Dataset

3.1 Motivation and Theoretical Framing

Existing emotion classification corpora such as *EmpatheticDialogues* (Rashkin et al., 2019) and *GoEmotions* (Demszky et al., 2020) have laid the foundational work in emotion classification, but focus exclusively on text, overlooking prosodic and embodied dimensions central to emotional communication such as tone, timing, and gesture. This unimodal framing limits progress towards cognitively grounded emotional reasoning.

Drawing from analogical reasoning (Gentner, 1983) and affective conceptual spaces (Gärdenfors, 2000), we argue that emotional intelligence in LLM requires a behaviorally informed context, not just surface level fluency. Our work addresses this by building on nonverbal signals to support more situated and expressive affective understanding in generative systems.

3.2 Emotion-Centric Interview Design

We structured MECC using a 60-item psychological questionnaire derived from affective science frameworks. The prompts elicited responses spanning 15 emotion categories:

- **Primary Emotions**: Love & Affection, Anger & Frustration, Fear & Anxiety, Happiness & Joy, Sadness & Grief, Guilt & Regret, Loneliness & Isolation
- **Self-Reflective Cognition**: Confidence & Self-Belief, Decision-Making, Forgiveness & Letting Go, Emotional Growth & Self-Reflection
- **Social-Affective Constructs**: Empathy & Understanding Others, Gratitude & Contentment, Stress & Coping, Non-Verbal Communication

Each prompt was either introspective or scenario-driven (e.g., “How do you process emotional failure?”, “What helps you remain resilient in uncertain times?”), encouraging participants to articulate both immediate and reflective affective states.

3.3 Interview-Based Data Collection and Annotation

We conducted semi-structured interviews with 31 participants (ages 18–35) based on a 60-item psychological questionnaire designed to elicit introspective responses across 15 emotional-cognitive categories. Each interview lasted approximately 45 minutes and was conducted in a quiet, controlled setting. During each session, the interviewer posed the questions while a second trained observer who was trained in both affective psychology and behavioural coding. Annotations captured three key behavioural dimensions:

Tone: calm, reflective, hesitant, defensive, etc.

Response Time: fast (1–2s), moderate (3–4s), or slow (5–6s)

Body Language: gestures, gaze, posture shifts, observed live during the session

The annotation protocol was designed to capture both deliberate and spontaneous behaviors.

3.4 Speech-to-text Transcription

Sessions were audio-recorded and audio recordings were transcribed using OpenAI's Whisper (large-v2), a s.o.t.a ASR system optimized for spontaneous and conversational speech. To ensure high quality transcriptions suitable for downstream emotion modeling we applied a multistage pipeline comprising:

- **Disfluency normalization**: Removal of filler words, repetition and hesitation while preserving emotional cues
- **Manual correction**: Rectification of misrecognized tokens and named entities
- **Quality Filtering**: Removal of short or emotionally uninformative responses

This pipeline preserved both semantic fidelity and temporal expressiveness elements vital for modeling affective grounding in language.

3.5 Emotion Labeling Strategy

Emotion labels were manually assigned during the annotation phase, before any model training. Rather than defaulting to the emotion implied by the question, each response was labeled based on the participant's expressive intent and affective tone. For example, if a participant responded with optimism and conviction to a question framed around fear, the label *Confidence & Self-Belief* was used instead of *Fear & Anxiety*. This response-grounded labeling aligns with contemporary research advocating for expression-based emotional classification, thereby improving downstream emotional alignment in generative models.

3.6 Data Representation and Structure

All annotated data was structured in a flattened JSON format suitable for both fine-tuning and metadata-aware retrieval in Retrieval-Augmented Generation (RAG) systems. Behavioural metadata, such as tone, response time, and body language, are embedded directly into the prompt. This structure enabled the model to condition not only on the semantic question but also on paralinguistic and behavioral framing, a design choice critical for emotionally coherent generation.

3.7 Dataset Statistics

Property	Value
Total Responses	*1764*
Unique Participants	31
Emotion Categories	15
Avg. Response Length	71.3 tokens
Avg. Response Time	3.4 seconds
Entries with Metadata	100%
Format	JSON

Table 2: Summary statistics of the constructed dataset.

4 Methodology

This work introduces a dual-path generation framework designed to simulate emotionally intelligent language by conditioning both learning and inference on behavioral context. Unlike traditional LLM pipelines that rely solely on semantic content, our method integrates paralinguistic metadata—such as tone, response time, and body language at each stage of training and generation. This methodological design supports cognitively and affectively aligned response generation across diverse emotional domains.

4.1 Overview

Our framework generates emotionally grounded responses by integrating metadata-filtered retrieval with an instruction-tuned large language model. We propose the dual-path architecture shown in Figure 2 , consisting of:

- A **fine-tuned LLaMA 3.1–8 B-Instruct model**, trained on behaviour-embedded prompts.
- A **Retrieval-Augmented Generation (RAG)** pipeline, retaining and encoding responses that retrieve semantically and affectively comparable exemplars from a metadata-indexed collection.

Both modules are designed to reinforce one another: while fine-tuning helps the model internalize behavioral priors, retrieval ensures each response remains grounded in real-world, affectively annotated human data. The overall objective is to ensure that generated responses exhibit both emotional fluency and contextual appropriateness.

4.2 Behavioral Prompt Construction

MECC includes structured (Q, A) interview pairs, annotated with multimodal behavioural metadata (

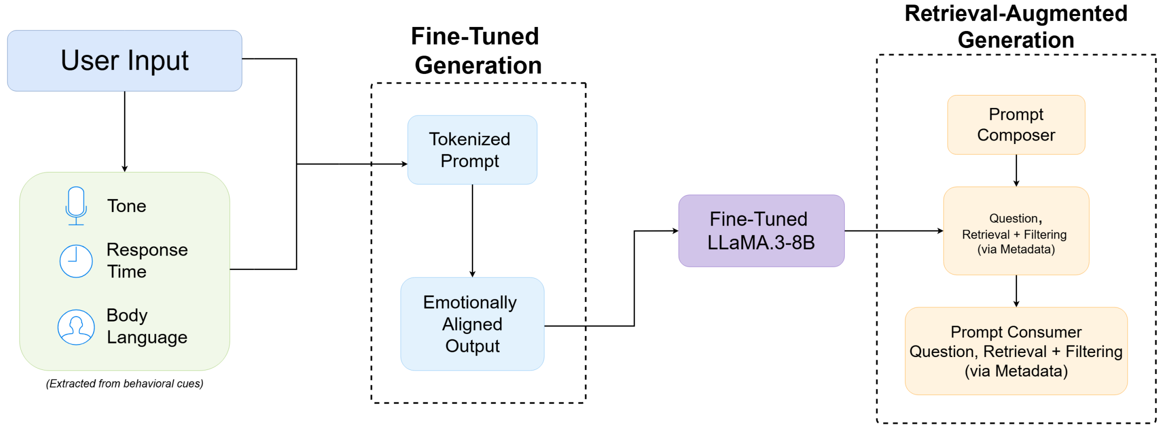

Figure 2: Dual-Path Emotionally Aware Generation Framework via Fine-Tuning and Analogical Retrieval-Augmented Generation

tone, response time, body language) and one of 15 target emotion categories. To support instruction-based learning, we flatten this structure into a JSON-style prompt, incorporating all behavioural and affective cues, suitable for both fine-tuning and retrieval tasks. This design ensures that the model learns from both semantic intent and behavioral expression, aligning generated responses with the emotional subtext of each question.

4.3 Emotionally Aware Language Modeling

We use the LLaMA 3.1–8 B-Instruct model due to its effective instruction-following capabilities. For scalable fine-tuning on consumer GPUs, we used LoRA with 8-bit NF4 quantization (via `bitsandbytes`), conserving memory while maintaining model fidelity.

The first generation path involves fine-tuning a LLaMA 3.1–8B-Instruct model on a behaviorally enriched dataset, enabling it to internalize patterns of emotional expression beyond surface-level language. This is achieved through Parameter-Efficient Fine-Tuning (PEFT) using the LoRA technique.

This design enables the model to generate responses that exhibit emotional intelligence, grounded in the behavioral and affective context of the prompt.

4.4 Emotionally Aligned Retrieval-Augmented Generation(RAG)

To enrich generation with affective grounding, we implement an Emotion-Conditioned Retrieval-Augmented Generation (RAG) mechanism. All training responses are indexed using `FAISS` (Facebook AI Similarity Search), an open-source library designed for efficient similarity search and clustering of dense vectors. Sentence-level embeddings are generated using the `all-MiniLM-L6-v2` model from SentenceTransformers. Each document in the index is annotated with behavioral metadata, including emotion label, tone, response time, and body language, enabling metadata-aware filtering during retrieval.

4.5 Inference Flow

At inference time, the system processes each user query through two complementary generation pathways:

Fine-Tuned Generation: The user's question, along with its associated behavioral metadata (tone, response time, body language), is directly passed to the fine-tuned `LLaMA 3.1-8B-Instruct` model. This enables emotionally aligned generation by conditioning the response on embedded affective signals.

RAG-Enhanced Generation: A metadata-indexed FAISS corpus is constructed from the MECC dataset, where each response is embedded using `all-MiniLM-L6-v2` and annotated with emotion, tone, response time, and body language. During inference, a joint scoring function identifies the top-k behaviorally congruent exemplars using semantic similarity (cosine distance), emotion alignment (e.g., `emotion=gratitude`), tonal and temporal congruence, and fingerprinting-based diversity filtering. These retrieved responses are concatenated into a context block and prepended to the user query, forming an augmented prompt. The final generation is then produced by the fine-tuned LLaMA model, grounded in both semantic and emotional context.

This dual-pathway architecture enables the system to simulate emotionally intelligent dialogue by internalizing affective behavior through fine-tuning, while dynamically adapting outputs using context-aware retrieval.

5 Experimentation

5.1 Experimental Setup

To evaluate our framework's potential for emotionally grounded generation, we performed experiments with the MECC dataset of behav- iorally annotated question-response pairs. We chose to examine two basic generation methods: (1) fine-tuned generation via LoRA; and, (2) emotion-driven RAG.

We completed fine-tuning the LLaMA 3.1–8B-In- struct model on instruction-style prompts, which contained behavioral metadata (tone, response time, body language) and naturally-formed language questions. Fine-tuning was completed using parameter-efficient fine-tuning (PEFT) - specifically, we applied Low-Rank Adapters (LoRA) to the attention projection layers (q_proj, k_proj, v_proj, o_proj) with a rank of 8, a scaling value (alpha) of 16, and a dropout of 0.1 which allowed us to balance regularization and performance. We trained the model for 3 epochs with a learning rate of 5e-5, a batch size of 2 per device, and 4 gradient accumulation steps; defining an total batch size of 8.

The emotionally grounded RAG component was implemented with FAISS for dense vector retrieval. Sentence embeddings were built with the all-MiniLM-L6-v2 model and indexed with metadata - emotion, tone, response time - for each response. For inference, the system conducted top-k retrieval with a single score from score of semantic similarity, emotional match, and behavioral concordance. The retrievals were concatenated and prepended to the user prompt to inform the model's final generation.

We ran our models and evaluated them using a test set of 353 samples, and the responses spanned 15 emotional categories, with each of the sample responses including behavioral metadata about response time, tone, and body language.

All of the model responses were generated with temperature between 0.4 and 0.7, 2048 token limits, and 1.1 repetition penalty to ensure coherence. Assessment of models focused on emotional alignment, semantic relevance, and behavioral fidelity from both generation paths.

5.2 Experimental Results

5.2.1 Performance Comparison: RAG vs. Non-RAG

- **Emotional Accuracy:** The non-RAG model yielded slightly higher emotional accuracy (39.94%) compared to the RAG model (38.24%).

- **Semantic Similarity:** The non-RAG model achieved a perfect BERTScore F1 (1.000), whereas the RAG model achieved 0.827. This suggests that while the non-RAG model may be overfitting or reproducing near-identical outputs, the RAG model is generating more varied and competitive responses.

- **Cosine Similarity:** Similarly, the non-RAG model scored a perfect cosine similarity (1.000), likely due to redundancy or training-test overlap. The RAG model's score of 0.443 indicates greater diversity in generated outputs.

- **Pearson Correlation & MSE:** Only reported for the RAG model. It achieved a Pearson correlation of 0.152 ($p = 0.004$) and a low mean squared error (0.091), demonstrating a mild but statistically significant alignment between predicted and ground-truth emotional intensities.

- **Perplexity:** The RAG model had a slightly higher perplexity (5.554) than the non-RAG model (4.386), suggesting more fluent and natural language generation.

Interpretation: While the non-RAG model performs slightly better on raw accuracy and textual similarity, this comes at the cost of overfitting and reduced response variability. The RAG model, though marginally lower in accuracy, delivers richer, more behaviorally grounded responses with statistically significant emotional alignment-indicating more human-like emotional reasoning. This supports our hypothesis that emotional intelligence cannot be evaluated solely through label accuracy but must consider generative diversity and affective coherence. Figure 3 and Figure 4 shows emotion distribution of targeted and predicted emotions.

Metric	RAG Model	Non-RAG Model
Emotional Accuracy	0.3824	0.3994
Average BERTScore F1	0.827	1.0000
Average Cosine Similarity	0.443	1.0000
Pearson Correlation	0.152 (p = 0.004)	0.3725 (p = 0.0000)
Mean Squared Error	0.091	15.78
Average Perplexity	5.554	4.386

Table 3: Comparison of performance metrics between RAG and Non-RAG models.

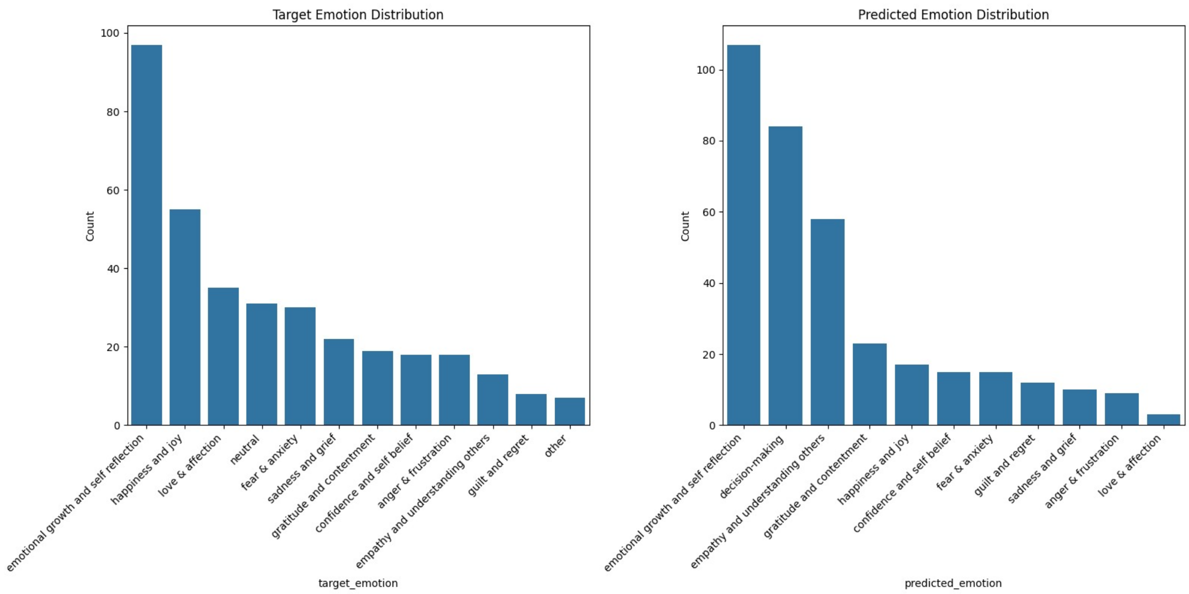

Figure 3: Target Emotion Distribution VS Predicted Emotion Distribution

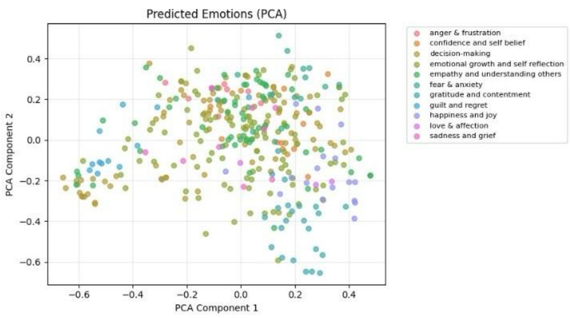

Figure 4: PCA-based clustering of predicted emotion distributions.

5.3 Qualitative Example of Emotional Grounding

A representative example is provided in Appendix , illustrating the model's ability to generate emotionally grounded responses. It shows how the system leverages metadata to produce outputs with improved emotional resonance and context sensitivity offering insights that complement the quantitative findings.

5.4 Emotion-Level Analysis

Emotion	F1-Score	Precision	Recall
Guilt and Regret	0.7000	0.5833	0.8750
Anger & Frustration	0.6667	1.0000	0.5000
Fear & Anxiety	0.6667	1.0000	0.5000
Confidence and Self Belief	0.5455	0.6000	0.5000
Gratitude and Contentment	0.5238	0.4783	0.5789
Emotional Growth and Self Reflection	0.5098	0.4860	0.5361
Sadness and Grief	0.5000	0.8000	0.3636
Happiness and Joy	0.4722	1.0000	0.3091
Empathy and Understanding Others	0.2817	0.1724	0.7692
Love & Affection	0.1579	1.0000	0.0857
Decision-Making	0.0000	0.0000	0.0000
Neutral	0.0000	0.0000	0.0000
Other	0.0000	0.0000	0.0000

Table 4: Per-emotion performance metrics showing F1-Score, Precision, and Recall for each emotion category.

6 Conclusion

This article has outlined a unified framework for generating emotionally and behaviorally informed responses in large language models. By fine-tuning a LLaMA-3.1 8B–Instruct model on MECC to enable the AI to generate affect-aligned responses based on more nuanced patterns of human cognition and emotion.

To optimize generative tuning, we have proposed a series of steps that constitute a metadata-filtered retrieval-augmented generation (RAG) pipeline that allows for ongoing dynamic semantically and emotionally matching responses with common affective metadata.

The framework also has two-track paths for coherence concerning emotional integrity, but also the contextual consistency of emotion. A series of validations along emotional alignment, semantic similarity measures, and regression measures demonstrated that our system was able to generate emotionally expressive and semantically well-grounded responses.

The proposed methodology demonstrates how combining multimodal fine-tuning with emotion-aware retrieval offers a scalable path toward constructing language models that are not only fluent but behaviorally grounded and emotionally intelligent.

7 Future Work

In addition to the possibilities outlined in our current framework, we suggest a multi-agent architecture for fine-grained emotion inference, balancing distributed specializing with orchestrated decision-making. This future architecture envisions 15 specialized emotion agents, each trained to detect a specific emotional state such as joy, fear, love, or coping with stress based on behaviorally grounded psychological patterns. Each agent is a separate and autonomous microservice, and each microservice can evaluate its inputs in parallel with various prompt strategies, each rooted in the focused emotional space.

An Emotion Orchestrator will premise and execute the coordination of these agents in four steps: 1) Emotion Probability Estimates through a question classifier 2) Task Distribution to selected emotion agents 3) Score Aggregation through the outputs from the emotion agents 4) Final Emotion Scoring through weighted fusing of the input from the respective agents.

In addition, we also wish to contribute a more robust dataset-level prompting framework that integrates pieces of compartmentalized persona traits, scene cues, and emotional framing—over the expected stimulus for each QA pair. These more sophisticated prompts will be key for both the classification step as well as the rationale of the emotion agents' decision-making process, since they will better ground the agent in a given context and priest its affective accuracy. By combining the complaint in specialization with the synergy of orchestration and persona-level prompting, we aim to create a framework that enhances the interpretability, scalability, and emotional fidelity of affect-aware

8 Limitations

While our proposed framework introduces a novel dual-path architecture for emotionally grounded generation, several limitations must be acknowledged.

- **Dataset Scale and Diversity:** The MECC dataset, while rich in multimodal annotations, is limited to 31 participants. This constrains the model's ability to generalize across diverse cultural, demographic, and communicative contexts.
- **Emotion Coverage Imbalance:** Certain emotional categories such as *Love & Affection* and *Neutral* were significantly underrepresented. This imbalance led to skewed performance across emotion classes and hindered the evaluation of low-frequency affective states.
- **RAG Output Variability:** Although the RAG pathway promotes semantic diversity, it occasionally retrieves contextually misaligned examples, introducing inconsistency in emotional fidelity and response tone.
- **Subjectivity in Annotation:** Despite trained annotators, the labeling of emotions and behavioral cues such as tone, body language, and response latency remains inherently subjective, which may introduce noise in alignment with ground-truth labels.
- **Lack of Human-Centric Evaluation:** The current evaluation relies primarily on automated metrics for emotion classification and semantic alignment. Human assessments of emotional appropriateness, empathy, and conversational relevance were not incorporated and remain a critical area for future work.

9 Ethics Statement

The final informed consent was granted by all persons participating in the data collection. In adherence to ethical principles of human subjects research, the interviewer was expected to have disclosed to the interviewees the study's objectives, data use, and their rights to withdraw from the study at any time. A pseudonym was maintained for all participants. All the audio recordings of the interviews had been transcribed and then anonymized to maintain the confidentiality of the participants.

Annotators were trained to identify behavioral and emotional markers, were instructed to remain as neutral as possible and be respectful during the labeling process, and were instructed in the management of all such data according to their respective institutions' policies and research guidelines. We hope that this work will aid in creating systems that have emotional intelligence while applying responsible AI principles related to transparency, dignity, and fairness.

References

Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024. Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359*. Accepted to ACL 2024 (Findings).

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. Main Conference Track.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*. Accepted to ACL 2020.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170. Available online 30 November 2004.

Peter Gärdenfors. 2000. *Conceptual Spaces: The Geometry of Thought*. The MIT Press. In Special Collection: CogNet.

Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *IVA '19: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 111–118.

Keith J. Holyoak and Paul Thagard. 1995. *Mental Leaps: Analogy in Creative Thought*. MIT Press. Paperback.

Kenneth Y T Lim, Minh Anh Nguyen Duc, and Minh Tuan Nguyen Thien. 2024. Exploring the potential of ai in nurturing learner empathy, prosocial values and environmental stewardship. *arXiv preprint arXiv:2408.15906*. 14 pages, 8 figures.

Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. 2024. Nus-emo at semeval-2024 task 3: Instruction-tuning llm for multimodal emotion-cause analysis in conversations. *arXiv preprint arXiv:2501.17261*. To appear in SemEval-2024 proceedings.

Albert Mehrabian. 1971. *Silent Messages*. Academia.edu. Cited by 5437.

Prashant Kumar Nag, Amit Bhagat, R. Vishnu Priya, and Deepak Kumar Khare. 2023. Emotional intelligence through artificial intelligence: Nlp and deep learning in the analysis of healthcare texts. In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*. IEEE. Presented at ICAIIHI 2023, Raipur, India. Added to IEEE Xplore on April 15, 2024.

E. Ortega-Ochoa, J.-M. Sabaté, M. Arguedas, J. Conesa, T. Daradoumis, and S. Caballé. 2024. Exploring the utilization and deficiencies of generative artificial intelligence in students' cognitive and emotional needs: A systematic mini-review. *Frontiers in Artificial Intelligence*, 7:1493566.

Rosalind W. Picard. 1997. *Affective Computing*. The MIT Press. Hardcover, 67.50.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. *arXiv preprint arXiv:1811.00207*. Accepted to ACL 2019 (long paper).

Sandeep Varma, Shivam Shivam, Sarun Natarajan, Biswarup Ray, Bagesh Kumar, and Om Dabral. 2024. Talk to your brain: Artificial personalized intelligence for emotionally adaptive ai interactions. In *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, Prayagraj, India. IEEE.

Fanfan Wang, Heqing Ma, Jianfei Yu, Rui Xia, and Erik Cambria. 2024. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. *arXiv preprint arXiv:2405.13049*. Accepted to the 18th International Workshop on Semantic Evaluation (SemEval-2024).

Hua Wang, Helmut Prendinger, and Takeo Igarashi. 2004. Communicating emotions in online chat using physiological sensors and animated text. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, pages 1171–1174. ACM.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

A Appendix

Prompt Template

```
<s>[INST]

<SYS>
Reply in a calm and reflective tone, reflecting the emotional state:
sadness and grief.
Behavioral cues:
- Response latency: slow
- Body language: gaze aversion, downward posture
</SYS>

How do you deal with personal loss?

[/INST]
I usually take time to sit with my feelings.  Sometimes I write in
a journal or talk to a close friend, but I try not to suppress the
sadness. It helps me heal.
</s>
```

Dataset structure

"input": "Q: What makes you feel most loved or appreciated in a relationship?
Tone: thoughtful, calm,
Response Time: Moderate (3–4 s),
Body Language: lowered gaze, soft voice"
"output": "I feel most loved when someone accepts me as I am completely."

LoRA Fine-Tuning Configuration

LoRA Rank	8
LoRA Alpha	16
Epochs	3
Batch Size	2 (gradient accumulation = 4)
Learning Rate	5×10^{-5}
Dropout	0.1
Precision	FP16 + gradient checkpointing
Optimizer	AdamW (max grad norm = 0.3)
Warmup Steps	10% of total steps
Target Modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`

Example Output rag v/s non-rag

Q: How do you typically cope with feelings of loneliness? [
“tone”: “optimistic, sincere”,
“response_time”: “Moderate (3-4s)”,
“body_language”: “gentle smile, relaxed shoulders”
]
Target: happiness and joy | Predicted: happiness and joy

RAG Response: When I feel lonely, I typically cope by engaging in activities that connect me with others and boost my mood. I might reach out to friends or family through video calls, which helps maintain meaningful relationships. I also find that pursuing hobbies like reading or listening to music creates a sense of purpose and joy. Additionally, I’ve learned that volunteering or helping others can transform loneliness into a feeling of contribution and connection to my community.

Non-RAG Response: I completely understand how isolating loneliness can feel. When I experience it, I find comfort in watching movies, especially comedies that genuinely make me laugh and remind me of brighter moments. There was a time when I first moved away from my family and felt so disconnected - those movies became my companions during difficult evenings. I also make sure to call my family regularly because hearing their voices instantly reminds me that I’m loved and that everything will be okay, even when distance makes it hard to remember.