Multimodal Empathetic Response Generation

Sean Gai
University of Southern California
Los Angeles, CA, United States
sgai@usc.edu

Zach Daniels
University of Southern California
Los Angeles, CA, United States
zsdaniel@usc.edu

Jiwon Hae
University of Southern California
Los Angeles, CA, United States
hae@usc.edu

ACM Reference Format:
Sean Gai, Zach Daniels, and Jiwon Hae. 2025. Multimodal Empathetic Response Generation. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM DEFINITION

Empathy in conversation arises from a complex interplay of text, tone, and facial expressions, making emotional understanding challenging for computational models. Effective empathetic response generation requires integrating multiple modalities – audio, vision, and text – to accurately interpret emotions.

A key challenge is fusing inputs from these diverse modalities. Emotionally and contextually-relevant response generation requires a nuanced understanding of conversational history and emotional progression. This project aims to develop a multimodal model that improves emotional understanding and generates coherent and empathetic responses.

2 RELATED WORK

2.1 Traits to Empathy: Personality-Aware Multimodal Empathetic Response Generation

From Traits to Empathy: Personality-Aware Multimodal Empathetic Response Generation [12] introduces a framework that integrates visual, audio and textual information along with speaker personality to enhance empathetic response generation. Unlike prior studies that relied solely on textual data, this approach leverages multimodal features and personality traits to model affective and cognitive characteristics of the speaker, leading to more nuanced and empathetic responses.

To validate their approach, the authors conducted experiments using MELD [8] and MEDIC [13], two multimodal datasets that provide video, audio and text.

The proposed framework consists of four main components: the refine encoder, fusion encoder, personality indicator, and empathetic response generator. It first processes textual and visual data using GPT-2 for text-based embeddings and BLIP [7] for vision-language representation. These representations serve as inputs to separate encoding modules that refine and align multimodal features.

Refine Encoder & Fusion Encoder
The refine encoder distills the visual features into refined visual features using query-key-value self-attention. Similarly, the fusion encoder processes textual representations through self-attention, producing an encoded matrix. It then uses the query matrix of the textual modality and the key matrix of the visual modality to perform cross-modal attention.

Personality Indicator
The framework incorporates a pre-trained MBTI personality indicator trained on Kaggle's MBTI dataset [9] to infer personality traits from textual inputs.

Empathetic Response Generator
The empathetic response generator aggregates utterances and control signals – emotions and personality traits – within a dialogue, using special tokens and embeddings to structure the input and generate responses.

All components are trained end-to-end, enabling the model to learn complex interactions among modalities and improve the empathetic quality of generated responses.

2.2 ImageBind

ImageBind is a model developed by Meta AI that learns a shared embedding space across six modalities – images, text, audio, depth, thermal, and IMU data [1]. The joint embedding space enables 'out-of-the-box' multimodal applications such as cross-modal retrieval, cross-modal embedding composition with arithmetic, and cross-modal generation.

Implementation-wise, the model consists of an encoder and linear projection head for each modality. The encoders for each modality are summarized in Table 1. The modality-specific embeddings produced by these encoders are subsequently mapped into the d-dimensional joint embedding space by the linear projection heads. The encoders and embeddings are trained to minimize InfoNCE contrastive loss on image-paired data. After aligning each modality's embeddings to image embeddings, the authors observed emergent alignment across all of the modalities, suggesting that the joint embedding space can effectively capture cross-modal relationships. Furthermore, the authors discovered that by applying embedding space arithmetic (e.g., addition), they could generate composed embeddings that effectively capture semantics from different modalities. For example, by adding the image embedding of a bowl of fruit

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Modality	Encoder
Images	Pre-trained ViT-H (630M) from OpenCLIP
Text	Text encoder (302M) from OpenCLIP
Audio	Convert to 2D mel-spectrogram, use ViT-B
Depth	Treat as one-channel image, use ViT-B
Thermal	Treat as one-channel image, use ViT-S
IMU	Extract X, Y, Z accelerometer & gyroscope measurements, project with 1D convolution, encode resulting sequence with Transformer

Table 1: Modality-specific encoders for ImageBind

and the audio embedding of chirping birds, they could perform image retrieval and obtain an image of a fruit tree.

Inspired by Personality-Aware Multimodal Empathetic Response Generation, our work also aims to fuse multimodal inputs to generate empathetic responses, but we intend to use ImageBind and early fusion to simplify the architecture.

3 DATASET

To explore our idea, we use MELD [8], one of the widely used multimodal datasets. MELD comprises over 1,400 dialogues and 13,000 utterances sampled from the television series Friends. Each utterance includes audio, visual, and textual modalities, and is annotated with one of seven emotions (anger, disgust, sadness, joy, neutral, surprise, or fear) and one of three sentiments (positive, negative, or neutral). The dataset features a diverse set of dialogues where multiple characters interact through speech and subtle emotional cues.

MELD's multimodal nature enables a more comprehensive understanding of affective expression, making it particularly well-suited for our study, where we aim to leverage multimodal information for generating empathetic responses. Table 2 provides a summary of MELD's key statistics.

	Train	Dev	Test
# of modality	{audio, visual, text}		
# of unique words	10,643	2,384	4,361
Avg. utterance length	8.03	7.99	8.28
Max. utterance length	69	37	45
Avg. # of emotions per dialogue	1039	114	280
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

Table 2: Summary of MELD's key statistics

4 METHODS

4.1 Data Preprocessing

Dialogue History
Each utterance in the dataset is uniquely identified by Dialogue_ID and Utterance_ID. To get the dialogue history up to and including each utterance, we concatenate Speaker and Utterance for all previous utterances within the same dialogue.

Audio Extraction
To get audio waveforms for each utterance, we use FFmpeg to extract the audio as a .wav file from the corresponding .mp4 of each utterance. Audio extraction fails for two utterances in the dataset, reducing the sizes of our training and development sets to 9,988 and 1,108, respectively.

Face Cropping
The videos in the MELD dataset often feature multiple participants or, in some cases, no visible participants. To focus on capturing the speaker's emotional expression, we crop videos to include only the speaker's face when visible and exclude samples where the speaker is not visible.

To perform face cropping, we employ active speaker detection to identify the speaker. TalkNet [11], introduced in 2021, represents the state-of-the art in active speaker detection. It incorporates audio and visual temporal encoders for feature representation, along with an audio-visual cross-attention mechanism to facilitate inter-modal interaction. Using the pretrained weights provided by the authors, we apply TalkNet to compute active speaker confidence scores, as shown in Figure 2. We then crop the face of the speaker with the highest confidence score to generate speaker-only videos as illustrated in Figure 3. After face cropping, the sizes of our training, development, and test sets were reduced to 9,648, 1,058, and 2,414, respectively.

Empathetic Target Response Generation
Since Friends is a sitcom, many responses are sarcastic, comedic, or nonsensical. To generate empathetic ground-truth responses for fine-tuning and evaluation, we leverage Gemini 2.0 Flash-Lite [2] using the prompt detailed in Listing 1.

4.2 Baselines

Pretrained LLMs (Text Only)
Our first two baselines leverage the pretrained Mistral-7B-Instruct-v0.3 [6] and Llama-3-8B-Instruct [3] models to process text-only dialogue history and generate responses *without fine-tuning*.

Fine-tuned LLMs (Text Only)
Our next set of baselines builds upon the pretrained models by fine-tuning them on our synthetic empathetic MELD dataset. We use Quantized Low-Rank Adaptation (QLoRA) [4] for fine-tuning due to GPU memory constraints.

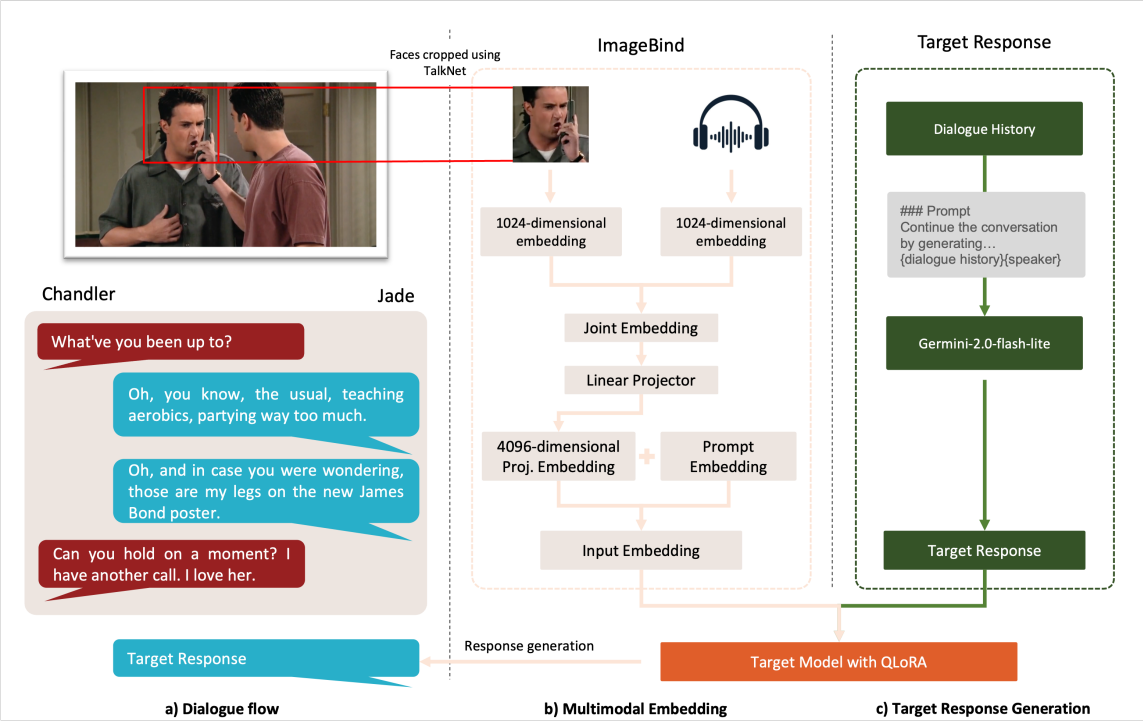

Figure 1: An illustration of the proposed multimodal empathetic response generation pipeline. a) depicts a sample dialogue, b) shows audio and visual preprocessing and fusion, and c) illustrates target response generation.

4.3 Imagebind MLLM

Our proposed model utilizes a multimodal architecture that integrates audio, video, and text inputs to generate empathetic responses as demonstrated in Figure 1. The audio and visual inputs from MELD are processed by ImageBind to get modality-specific embeddings, which are then added element-wise to create a joint embedding. The joint audio-visual embedding is then mapped to the LLM's embedding space via a Linear Projector. Next, the projected audio-visual embedding is prepended to the prompt embeddings taken from Listing 1, which include the instructions and dialogue history, to get the full input embeddings. Finally, the full input embeddings are used to generate responses from the LLM.

The weights of the linear projector are all made trainable, while ImageBind remains frozen to minimize the number of trainable parameters.

Additionally, instead of fine-tuning the entire language model, we adopt a parameter-efficient training strategy by combining 4-bit quantization with LoRA. More specifically, we used 4-bit weight loading with double quantization and float16 computation. LoRA is configured with a rank of $r = 8$, scaling factor $\alpha = 16$, and dropout rate of 0.1.

5 RESULTS

We evaluated the empathetic response generation capabilities of our baselines and proposed multimodal approach across four criteria: fluency, diversity, semantic similarity, and polarity. The results are summarized in Table 3.

5.1 Polarity Scoring

To assess the empathetic quality of generated responses, we compute three polarity metrics: *emotion*, *sentiment*, and *empathy*. These metrics evaluate how well a response aligns with the emotional and sentiment context of the preceding dialogue.

The *emotion* metric is calculated by using the bhadresh-savani/distilbert-base-uncased-emotion [10] model to generate soft emotion distributions for each utterance in the dialogue history and for the response itself. The dialogue history's distributions are averaged, and cosine similarity is used to compare this average with the response's distribution. This yields a score between 0 and 1 to indicate emotional alignment.

Methods	Model	Fluency	Diversity		Semantic Similarity (BERTScore)			Polarity		
		PPL	Dist-1	Dist-2	Precision	Recall	F1	Emotion	Sentiment	Empathy
Baseline (text only)	Mistral-7B-Instruct-v0.3	23.56	**0.0982**	**0.3110**	0.8756	0.8778	0.8765	0.58	**0.69**	**0.63**
	Llama-3-8B-Instruct	34.18	0.0676	0.2627	0.8539	**0.9572**	**0.9022**	0.64	0.72	0.68
Fine-tuned Baseline (text only)	Mistral-7B-Instruct-v0.3	**17.82**	0.0684	0.1965	**0.8907**	0.8879	0.8891	0.56	0.68	0.62
	Llama-3-8B-Instruct	**3.843**	0.0343	0.1585	0.7986	0.9456	0.8656	**0.67**	**0.73**	**0.70**
Empathetic MLLM (Ours) (text + audio + video)	Mistral-7B-Instruct-v0.3	22.35	0.0914	0.2907	0.8883	**0.8961**	**0.8919**	**0.59**	0.67	**0.63**
	Llama-3-8B-Instruct	28.91	**0.0940**	**0.3260**	**0.8699**	0.8942	0.8817	0.54	0.64	0.59

Table 3: Evaluation of our method and the baselines on fluency, diversity, and semantic similarity metrics. Bold red figures indicate the best scores for Mistral-based models, while bold blue figures indicate the best scores for Llama-based models

The *sentiment* metric uses VADER [5] to compute compound sentiment scores (ranging from -1 to +1) for both the dialogue history and the response. The match score is defined as the inverse of their absolute difference:

$$\text{Sentiment Match Score} = 1 - |s_{history} - s_{response}|$$

indicating how closely the response's sentiment matches the overall sentiment of the context.

The *empathy* metric is the mean of the emotion and sentiment scores:

$$\text{Empathy} = \frac{1}{2}(\text{Emotion} + \text{Sentiment})$$

This combined metric captures both emotional coherence and sentiment alignment, providing a holistic measure of empathetic response quality.

6 DISCUSSION

As shown in Table 3, we evaluated our proposed Empathetic MLLM against both pretrained and fine-tuned text-only baselines across fluency, diversity, semantic similarity, and polarity. Fine-tuning the text-only models significantly improved fluency, with fine-tuned Llama-3-8B achieving a perplexity of 3.84 compared to 34.18 in its pretrained configuration.

Although our multimodal model did not reach the same fluency levels, it maintained competitive performance, particularly with the Mistral variant. One trade-off of the fine-tuned text-only models was a noticeable drop in diversity. This is likely due to overfitting, which led to repetitive and less varied responses. In contrast, our Empathetic MLLM models retained some of that lost diversity. The Llama-based MLLM achieved the highest Dist-2 score (0.3260), suggesting that adding audio and visual modalities maintained richer lexical variation.

Semantic similarity, as measured by BERTScore, remained high across all models, including the baselines, indicating strong alignment in meaning with the target responses. Notably, the Mistral-based MLLM achieved the highest BERTScore Recall and F1 among all Mistral models. Meanwhile, the Llama-based MLLM outperformed both text-only Llama baselines in terms of BERTScore Precision.

The polarity results provide further insights, particularly when examining the emotion scores. While the text-only fine-tuned Llama model achieved the highest empathy score of 0.70, our multimodal models performed slightly worse, especially on the emotion polarity metric. The Mistral and multimodal models exhibited lower emotion scores, which were skewed by a cluster of examples with scores near 0. These low-scoring cases generally fell into three categories. First, when the dialogue history was long and had high emotional entropy, it was difficult to generate a response that matched all emotional cues, often resulting in emotional mismatches and low scores. Second, short or plain generated responses lacked emotional content entirely, leading to neutral or misaligned emotion classifications. Third, some examples contained similar but distinct emotions—such as anger versus sadness—between the history and response, which led to mismatched predictions despite partially aligned emotional intent. When filtering out these challenging cases, the remaining emotion scores followed trends similar to the sentiment polarity scores, which suggests our models do learn to express emotion, even if inconsistently. See Figures 4 and 5 for a breakdown of polarity score distributions.

These findings highlight some limitations of our current approach, but also provide promising evidence that multimodal inputs can contribute to generating empathetic responses.

7 CONCLUSION

In this work, we proposed a multimodal approach to empathetic response generation by incorporating ImageBind to fuse audio, visual and text signals. Using our synthetic MELD dataset, we were able to improve certain BERTScore metrics using the Empathetic MLLM approach while reducing perplexity and maintaining similar levels of diversity and polarity alignment compared to the unimodal baselines.

Ideas for future work include incorporating human evaluation – particularly important given the difficulty of quantifying empathy with automated metrics – and expanding to additional datasets to support broader research in a variety of domains.

8 CONTRIBUTION

Name	Contribution
All	CARC setup Literature review Evaluation Report
Sean G.	Audio extraction MLLM implementation Mistral MLLM fine-tuning
Zach D.	Target response generation Mistral pretrained and finetuned baselines Polarity evaluation
Jiwon H.	Face cropping Llama pretrained and finetuned baselines Llama MLLM fine-tuning

REFERENCES

[1] Rohit Girdhar et al. "ImageBind: One Embedding Space To Bind Them All". In: *CVPR*. 2023.

[2] Google Cloud. *Gemini 2.0 Flash-Lite | Generative AI on Vertex AI*. 2024. URL: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash-lite.

[3] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[4] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[5] C.J. Hutto and Eric Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI, June 2014.

[6] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[7] Junnan Li et al. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". In: *ICML*. 2022.

[8] Soujanya Poria et al. "MELD: A multimodal multi-party dataset for emotion recognition in conversations". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 527–536.

[9] Gregorius Ryan, Pricillia Katarina, and Derwin Suhartono. "MBTI Personality Prediction Using Machine Learning and SMOTE for Balancing Data Based on Statement Sentences". In: *Information* 14.4 (2023).

[10] Bhadresh Savani. *distilbert-base-uncased-emotion*. https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion. Accessed: 2025-05-05. 2020.

[11] Ruijie Tao et al. "Is Someone Speaking?: Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection". In: *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21. ACM, Oct. 2021, pp. 3927–3935. DOI: 10.1145/3474085.3475587. URL: http://dx.doi.org/10.1145/3474085.3475587.

[12] Jiaqiang Wu et al. "From Traits to Empathy: Personality-Aware Multimodal Empathetic Response Generation". In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by Owen Rambow et al. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 8925–8938. URL: https://aclanthology.org/2025.coling-main.598/.

[13] Zhou'an Zhu et al. *MEDIC: A Multimodal Empathy Dataset in Counseling*. 2023. arXiv: 2305.02842 [cs.CV]. URL: https://arxiv.org/abs/2305.02842.

A TALKNET

Figure 2: Video with the speaker detected by TalkNet, with confidence scores displayed on top of the bounding box.

Figure 3: Video output after Face Crop

B EMPATHETIC RESPONSE GENERATION PROMPT

```
### INSTRUCTIONS ###
Continue the conversation by generating **only the next
     line** spoken by the indicated character.
Your response must be empathetic, showing understanding or
     emotional attunement to the preceding dialogue.

### EXAMPLE ###

=== DIALOGUE HISTORY ===
Rachel: Hey!
Ross: Hi!
Rachel: What are you doing here?
Ross: Ah y'know, this building is on my paper route so I...
Rachel: Oh.
Ross: Hi.
Rachel: Hi.
Ross: How'd did it go?
Rachel: Oh well, the woman I interviewed with was pretty
     tough, but y'know thank God Mark coached me, because
     once I started talking about the fall line, she got
     all happy and wouldn't shut up.
Ross:

=== RESPONSE ===
That sounds like a huge relief.

### TASK ###

=== DIALOGUE HISTORY ===
{Dialogue History}
{Next Speaker}:

=== RESPONSE ===
```

Listing 1: Prompt used for empathetic response generation

C POLARITY EVALUATION

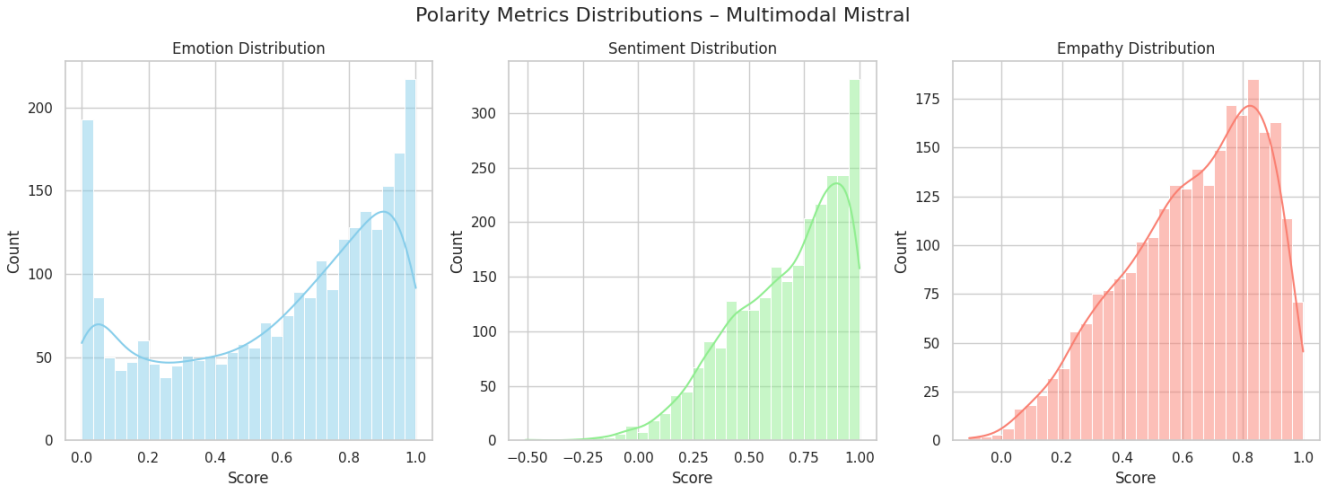

Figure 4: Mistral MLLM Polarity

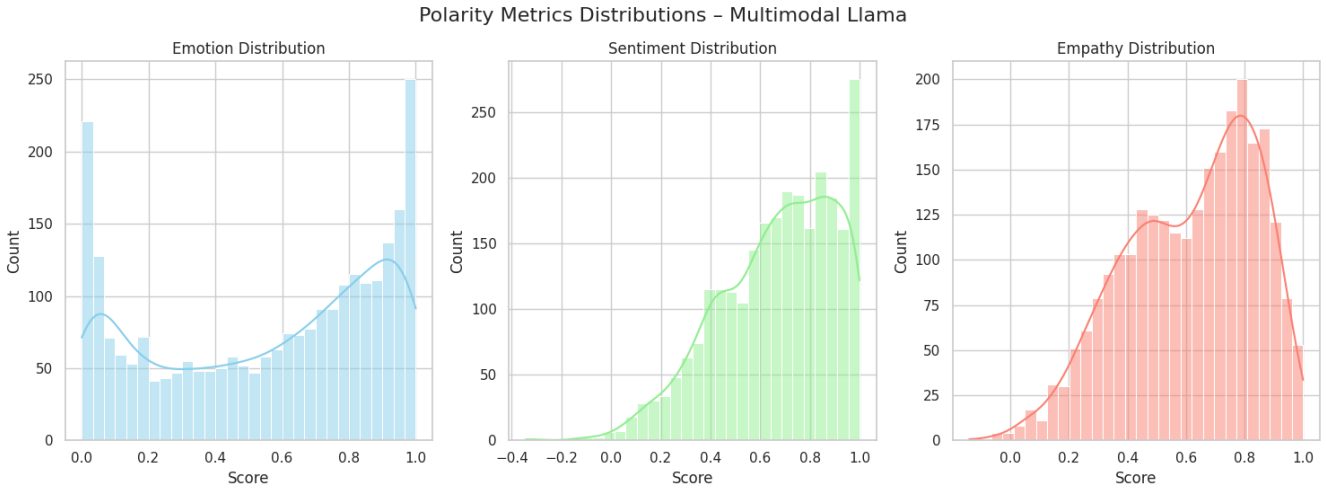

Figure 5: Llama MLLM Polarity