Analyzing Key Factors Influencing Emotion Prediction Performance of VLLMs in Conversational Contexts

Jaewook Lee[*], **Yeajin Jang**[*], **Hongjin Kim**, **Woojin Lee**, **Harksoo Kim**[†]
Konkuk University
{benecia428, dpwls258, jin3430, shes100, nlpdrkim}@konkuk.ac.kr

Abstract

Emotional intelligence (EI) in artificial intelligence (AI), which refers to the ability of an AI to understand and respond appropriately to human emotions, has emerged as a crucial research topic. Recent studies have shown that large language models (LLMs) and vision large language models (VLLMs) possess EI and the ability to understand emotional stimuli in the form of text and images, respectively. However, factors influencing the emotion prediction performance of VLLMs in real-world conversational contexts have not been sufficiently explored. This study aims to analyze the key elements affecting the emotion prediction performance of VLLMs in conversational contexts systematically. To achieve this, we reconstructed the MELD dataset, which is based on the popular TV series *Friends*, and conducted experiments through three sub-tasks: overall emotion tone prediction, character emotion prediction, and contextually appropriate emotion expression selection. We evaluated the performance differences based on various model architectures (e.g., image encoders, modality alignment, and LLMs) and image scopes (e.g., entire scene, person, and facial expression). In addition, we investigated the impact of providing persona information on the emotion prediction performance of the models and analyzed how personality traits and speaking styles influenced the emotion prediction process. We conducted an in-depth analysis of the impact of various other factors, such as gender and regional biases, on the emotion prediction performance of VLLMs. The results revealed that these factors significantly influenced the model performance.

1 Introduction

Emotional intelligence (EI) in artificial intelligence (AI), which refers to the ability of an AI to understand and respond appropriately to human emotions, is a crucial topic in AI research. EI involves the ability to interpret and manage the emotions that are embedded in information and is essential for various cognitive tasks, from problem solving to behavior regulation (Salovey et al., 2009). Human emotions play a significant role in various domains such as academics, competitive sports, and daily life and are shaped by internal and external factors (Koole, 2010; Pekrun et al., 2002; Lazarus, 2000; Li et al., 2023b). Equipping AI systems with EI enhances the quality of human-AI interactions, improves user experience, and enables more natural and effective communication based on emotional empathy.

Large language models (LLMs), which are considered a crucial step towards achieving artificial general intelligence, have exhibited exceptional performance in various fields (Bubeck et al., 2023). As a result, there has been growing interest in whether LLMs possess EI. Wang et al. (2023) evaluated the EI of LLMs through psychological measurements and discovered that GPT-4 achieved high EQ scores. Moreover, studies by Li et al. (2023b) and Li et al. (2023c) showed that LLMs can understand emotional stimuli in the form of text and images, perceiving emotions in a manner similar to humans. However, these studies have limitations as they focus on a single modality, whereas various factors such as verbal cues, visual cues, and contextual information interact in a complex manner in real-world conversational situations.

Vision large language models (VLLMs) have recently gained attention to overcome the above limitations. As VLLMs can process text and images simultaneously, they have the potential to solve more complex and multifaceted emotion prediction tasks. For example, VLLMs can infer emotional states by comprehensively analyzing the facial expressions and verbal cues of conversation participants or predict appropriate emotional responses considering

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5801–5816

the conversational context. However, despite their potential, the key factors influencing the emotion prediction of VLLMs in conversational situations have not yet been sufficiently explored.

This study aimed to analyze the factors influencing the emotion prediction of VLLMs, such as the model architecture, persona information, and biases, systematically to explore means of improving emotion prediction performance in conversational situations. To achieve this, we reconstructed the MELD dataset (Poria et al., 2018) based on the popular TV series *Friends* and augmented it by integrating various elements, such as images, conversational context, and persona information, to evaluate the performance of VLLMs comprehensively. We conducted an extensive assessment of the emotion understanding and expression performance of VLLMs through three sub-tasks: overall emotion tone prediction, character emotion prediction, and contextually appropriate emotion expression selection.

The experimental results showed that differences in the model architecture had a distinct impact on the emotion prediction performance. This suggests that the structural characteristics of VLLMs, such as the method of integrating image and text information and the LLM Backbone, play crucial roles in emotion prediction performance. In addition, models that included persona information exhibited notable differences in the emotion prediction process. This implies that information on the personality traits and speaking styles of an individual significantly influences the emotion understanding and response performance of the model. We also conducted an in-depth analysis of the impact of various factors related to emotion prediction, such as gender and regional biases, on the emotion prediction performance of VLLMs. The analysis revealed that factors such as gender and regional biases significantly influenced the emotion prediction process of VLLMs, revealing the biases and limitations that may arise in this process.

2 Related Work

The rapid development of LLMs has led to substantial progress in language generation, knowledge utilization, and complex reasoning tasks. However, as these models are being integrated into various application domains, enhancing their EI and mitigating social biases have become increasingly important. Wang et al. (2023) explored the EI of LLMs using psychological methods, thereby laying the foundation for further research on how these models perceive and respond to emotional stimuli. Building on this work, Li et al. (2023b) and Li et al. (2023c) investigated the ability of LLMs to understand emotional content and demonstrated that current models can react to emotional stimuli similarly to humans. Paech (2023) introduced a new criterion for evaluating the EI of LLMs through EQ-Bench, which is a benchmark that measures the ability of a model to predict the emotional states of characters within conversations. Sabour et al. (2024) proposed EMOBENCH, which is a benchmark that is designed to evaluate the EI of LLMs comprehensively by assessing not only emotion recognition, but also emotional regulation and the application of emotional understanding.

Along with research on the EI of LLMs, addressing the social biases that are inherent in these models is a crucial task for the development of ethical AI. Sheng et al. (2019) and Schick et al. (2021) emphasized the importance of recognizing and mitigating gender stereotypes and other biases in the training data. Nadeem et al. (2021) measured stereotypical biases using the StereoSet benchmark, while Parrish et al. (2022) evaluated biases in question-answering tasks using the BBQ dataset. These studies provided important insights into the EI and social biases of LLMs. Building on this foundation, the present study aimed to analyze the key factors influencing the emotion prediction of VLLMs in conversational contexts systematically. Specifically, we intended to investigate the impact of factors such as persona, gender, and regional biases on the emotion prediction processes of VLLMs in depth.

3 Dataset and Task Overview

We reconstructed the MELD dataset (Poria et al., 2018), which is based on popular TV series *Friends*, to investigate the key factors influencing the emotion prediction performance of VLLMs in conversational contexts. The MELD dataset provides full-scene images for each scene and the corresponding conversational context, along with the names of the characters who engage in the dialogue and the emotion and sentiment labels for the feelings of each character. The dataset includes emotion and sentiment labels for each utterance. Emotions are categorized into seven types: "fear," "disgust," "joy," "sadness," "surprise," "anger," and

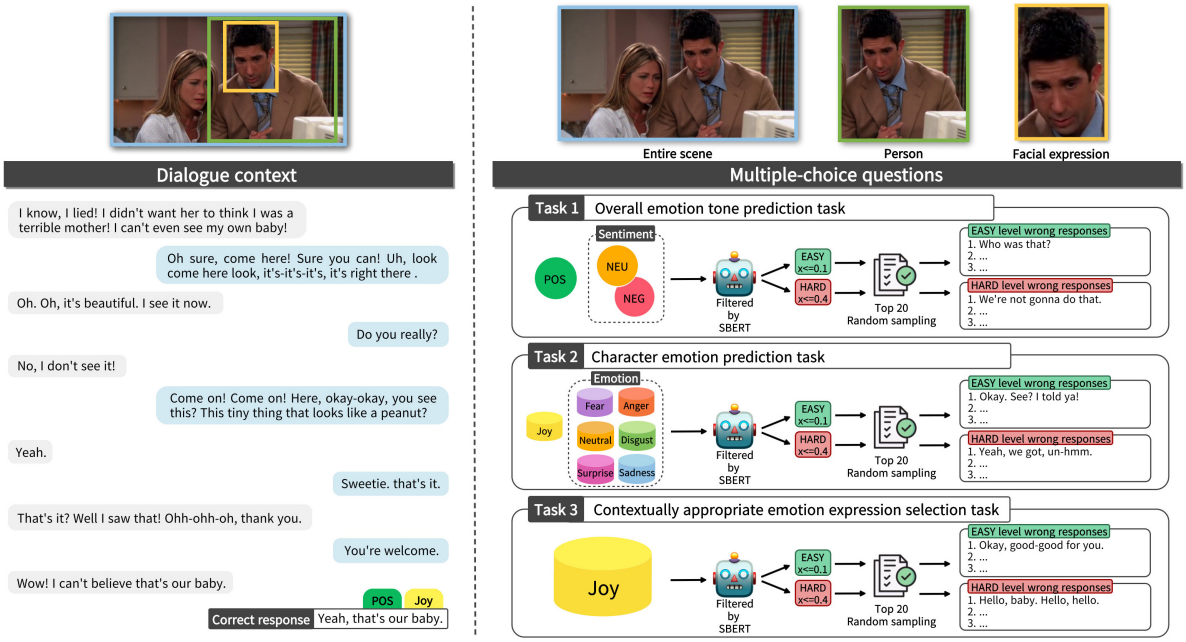

Figure 1: Overview of the data reconstruction process for evaluating the emotion prediction performance of VLLMs using the MELD dataset. The process involved three main stages: (1) dialogue selection, which filtered and adjusted dialogues based on the number of turns and presence of characters with personal information; (2) image scope reconstruction, which extracted images from video frames and categorized them into three scopes (entire scene, person, and facial expression) to capture different aspects of emotional information; and (3) incorrect sentence selection, which selected distractor sentences for each sub-task using SBERT.

"neutral," whereas sentiments are divided into three categories: "positive," "negative," and "neutral." Appendix A presents the overall statistics.

3.1 Persona Information

As the characteristics of an individual greatly influence emotion expression and understanding, we constructed additional persona information for the MELD dataset. The persona information consisted of the personality traits and speaking styles of each character.

Personality traits influence how individuals perceive and express emotions, and play a crucial role in understanding and modeling emotional responses in conversational contexts. We carefully defined the personality traits of the characters of *Friends* to provide comprehensive persona information. By including these personality traits in the model, we could investigate their impact on the emotion prediction performance.

Speaking styles affect how individuals convey their emotions and intentions. Each character of *Friends* has a unique manner of speaking. By integrating these speaking styles into the model, we could analyze their influence on emotion prediction performance.

3.2 Quantitative Evaluation

To evaluate the emotion prediction performance of VLLMs comprehensively, we approached the problem by selecting the most appropriate emotion expression in each conversational turn, beyond simply recognizing the emotions of the speaker. We used a multiple-choice question format in which each question consisted of one correct utterance and three incorrect utterances. The three subtasks were designed to assess different aspects of the emotion understanding and expression abilities of the model, as follows:

Overall emotion tone prediction task assessed the ability of the model to predict the overall emotional tone of the dialogue by selecting the most appropriate utterance from the options with different sentiments.

Character emotion prediction task evaluated the ability of the model to predict the emotions of specific characters in a given context by selecting the most appropriate utterance from the options expressing different emotions.

Contextually appropriate emotion expression selection task assessed the ability of the model to understand the context in depth and to select the

most appropriate emotion expression by identifying the correct utterance from options with the same emotion but different expressions.

4 Dataset Construction

We reconstructed the existing MELD dataset to align it with the objectives of this study. The data reconstruction process consisted of three stages: 1) dialogue selection, 2) image scope reconstruction, and 3) incorrect sentence selection. Using these stages, we constructed data that fit the purpose and removed unnecessary data. The entire process is illustrated in Figure 1.

4.1 Dialogue Selection

The original dataset includes various characters and sentences that are commonly used in real-life conversations. These characteristics are useful for identifying the key elements that influence the emotion prediction of VLLMs in conversational contexts. For data selection, we removed samples that either included dialogue with very long turns or could not reflect persona information.

Adjusting dialogues with very long turns. In conversations, instances arise in which very long turns appear. In such situations, models generally rely more heavily on the previous conversational context than on facial expressions or gestures. This can significantly affect emotion prediction, particularly for VLLMs that use LLMs as their backbone models. This is because these models may prioritize the textual context over visual cues. This can act as noise when identifying the key factors that influence the emotion prediction of the model.

Therefore, we decided to reduce dialogues exceeding 15 turns randomly, to between 9 and 15 turns. The reason for randomly adjusting the number of turns rather than fixing them was to prevent bias associated with the number of turns. In addition, the randomization ensured the inclusion of samples with various dialogue lengths within the dataset to aid in evaluating the model performance in real-life conversational scenarios with varying lengths.

Removing characters lacking persona information. We also aimed to evaluate the emotion prediction performance of VLLMs based on the inclusion or exclusion of persona information. To this end, we structured the dialogue data such that characters to whom persona information could be assigned appeared during the final utterance turn. However, collecting persona information for some characters (e.g., hosts, customers, and airline employees), is difficult or impossible. Therefore, dialogues involving such characteristics were excluded from the dataset. The final dataset included a pool of characters consisting of six main characters with persona information and 18 surrounding characters.

4.2 Image Scope Reconstruction

Text-based information is often effective for explicit communication, but has limitations in conveying complex emotional states or atmospheres. In contrast, images enrich these emotional nuances through nonverbal elements and visual context. Particularly in human conversations, emotions vary significantly depending on the context and environment. Therefore, the visual information contained in images, such as the posture, facial expressions, and gestures of the conversation partners, can capture the subtleties of emotions that are difficult to discern from text alone.

For image processing, the original videos were divided into frames and image information was extracted from each relevant frame. The most suitable frame was selected and used for the entire scene image. Person and facial expression images were extracted separately from the selected image, and the entire process was performed manually by the authors. At the end of each stage, cross-validation was performed to improve the image accuracy and ensure strict quality control.

4.3 Incorrect Sentence Selection

The final stage of the data construction involved the selection of incorrect sentences for each dialogue. In this stage, we selected incorrect sentences corresponding to the multiple-choice questions. We selected sentences with sentiments or emotions that differed from the correct sentences for the overall emotion tone and character emotion prediction tasks. For the contextually appropriate emotional expression selection task, we selected sentences with the same emotion as the correct sentence.

The selected sentences were filtered using SBERT (Reimers and Gurevych, 2019). Some sentences may have high semantic similarity and can be used interchangeably with the correct sentence; therefore, we removed sentences that received semantic similarity scores above a certain level to eliminate such cases. In addition, we constructed the dataset with two difficulty levels (easy and

Model	LLM	Tone			Emotion			Context			Avg.
		All	Person	Face	All	Person	Face	All	Person	Face	
Prompt type						*Original*					
InstructBLIP	Vicuna (13B)	40.23	40.15	39.10	40.28	40.83	40.45	40.72	41.09	41.16	40.45
LLaVA-1.5	Vicuna (13B)	50.39	50.15	49.77	48.98	48.94	48.75	48.69	48.39	47.50	49.06
LLaVA-Next	Vicuna (13B)	52.00	51.42	51.28	49.87	49.18	48.56	49.97	49.76	49.13	50.13
InstructBLIP	FLAN (11B)	56.10	56.36	56.59	56.75	56.75	56.92	55.35	55.34	55.73	56.20
LLaVA-1.5	Vicuna (7B)	38.11	38.41	37.45	36.49	36.63	36.67	35.72	35.48	34.82	36.65
LLaVA-Next	Vicuna (7B)	46.41	46.06	46.11	45.82	45.76	45.51	45.16	44.98	44.53	45.59
LLaVA-Next	Mistral (7B)	47.86	47.58	47.48	46.64	46.47	46.53	46.50	46.03	45.79	46.76
Qwen-VL-Chat	Qwen (7B)	39.78	39.49	40.17	38.88	38.92	38.52	37.75	37.97	37.56	37.86
MiniGPT-4	Vicuna (7B)	27.58	27.89	28.57	27.82	27.85	27.22	26.45	26.77	27.07	27.47
Otter	MPT (7B)	38.00	37.65	37.92	38.58	38.37	38.89	37.11	37.23	37.00	38.78
InstructBLIP	FLAN (3B)	51.91	51.91	51.37	51.86	51.57	51.52	50.75	50.64	50.06	51.28
Prompt type						*Personality traits*					
InstructBLIP	Vicuna (13B)	39.95	39.86	39.00	40.41	40.27	39.57	39.80	39.59	39.14	39.73
LLaVA-1.5	Vicuna (13B)	51.09	51.00	50.50	49.33	49.60	49.27	49.44	49.75	48.59	49.84
LLaVA-Next	Vicuna (13B)	49.94	49.05	48.55	47.23	46.73	46.33	48.55	48.47	47.28	48.02
InstructBLIP	FLAN (11B)	54.87	55.00	54.20	54.95	54.95	54.54	53.05	53.05	53.45	54.23
LLaVA-1.5	Vicuna (7B)	37.47	37.70	37.40	35.78	36.00	36.33	35.54	35.24	35.39	36.31
LLaVA-Next	Vicuna (7B)	44.98	44.92	44.70	44.89	44.42	44.83	44.92	44.98	43.98	44.73
LLaVA-Next	Mistral (7B)	46.27	46.08	46.22	45.53	45.51	45.22	45.13	45.33	45.13	45.60
Qwen-VL-Chat	Qwen (7B)	39.95	40.09	40.03	38.43	38.58	38.39	37.64	37.65	38.04	38.75
MiniGPT-4	Vicuna (7B)	28.15	29.09	29.32	28.39	28.51	29.05	28.66	29.11	28.66	28.77
Otter	MPT (7B)	38.76	38.78	38.80	39.34	39.31	39.45	37.88	37.88	38.33	38.73
InstructBLIP	FLAN (3B)	49.74	49.72	49.34	49.11	49.02	48.59	48.61	48.28	48.22	48.95
Prompt type						*Speaking styles*					
InstructBLIP	Vicuna (13B)	40.18	40.12	39.46	40.33	40.69	39.59	39.62	39.59	39.49	39.90
LLaVA-1.5	Vicuna (13B)	50.48	50.24	49.34	49.17	49.49	48.79	48.77	48.91	47.84	49.23
LLaVA-Next	Vicuna (13B)	50.45	49.66	49.67	46.87	46.95	46.23	47.47	47.00	47.12	47.94
InstructBLIP	FLAN (11B)	55.71	56.50	56.13	55.38	55.91	55.95	54.60	54.47	54.93	55.51
LLaVA-1.5	Vicuna (7B)	38.78	38.58	38.08	38.07	37.45	37.71	36.29	36.37	36.03	37.48
LLaVA-Next	Vicuna (7B)	45.13	45.48	45.08	44.56	44.65	44.97	44.81	44.67	44.59	44.88
LLaVA-Next	Mistral (7B)	47.88	47.78	47.20	46.57	46.47	46.03	46.12	46.27	45.91	46.69
Qwen-VL-Chat	Qwen (7B)	39.95	40.42	40.33	39.05	39.04	38.72	38.43	38.58	38.53	39.23
MiniGPT-4	Vicuna (7B)	29.12	29.30	29.11	29.38	29.75	30.05	29.39	29.65	29.59	29.48
Otter	MPT (7B)	39.34	38.97	39.45	39.44	38.97	39.58	38.10	38.33	38.69	38.98
InstructBLIP	FLAN (3B)	49.47	49.85	49.28	49.31	49.19	48.78	48.08	48.45	47.93	48.93
Prompt type						*CoT*					
InstructBLIP	Vicuna (13B)	40.53	40.47	39.17	41.52	41.02	40.24	41.17	41.19	40.41	40.63
LLaVA-1.5	Vicuna (13B)	49.33	49.04	48.94	49.15	49.13	49.06	48.13	47.94	47.22	48.66
LLaVA-Next	Vicuna (13B)	51.21	50.41	49.70	48.36	47.71	47.12	49.87	49.60	48.48	49.16
InstructBLIP	FLAN (11B)	55.94	56.11	56.02	55.84	55.94	56.17	54.93	55.06	55.05	55.67
LLaVA-1.5	Vicuna (7B)	41.88	42.16	41.50	41.44	41.33	41.08	39.62	39.69	38.39	40.78
LLaVA-Next	Vicuna (7B)	46.21	46.05	45.55	45.61	45.33	44.55	44.28	44.75	43.98	45.14
LLaVA-Next	Mistral (7B)	48.50	47.89	47.33	47.25	47.34	46.54	47.66	46.55	45.44	47.17
Qwen-VL-Chat	Qwen (7B)	39.54	39.24	39.45	37.50	37.35	38.05	36.86	36.64	36.31	37.89
MiniGPT-4	Vicuna (7B)	27.90	28.00	27.94	27.32	27.31	26.83	26.39	26.17	25.91	27.09
Otter	MPT (7B)	37.86	37.16	37.12	37.56	37.19	37.38	36.28	36.66	36.20	37.05
InstructBLIP	FLAN (3B)	52.14	52.01	51.95	52.29	52.05	51.90	50.95	50.93	50.48	51.64

Table 1: Comprehensive performance comparison of VLLM models with varied prompt types (original, personality traits, speaking styles, and CoT). The results, shown as the accuracy scores averaged across three distinct prompts per type, indicate the mean performance on the easy and hard levels. "All" denotes entire scene scope, "Person" refers to individual character scope, and "Face" refers to facial expression scope. "Tone," "Emotion," and "Context" correspond to the overall emotion tone prediction task, character emotion prediction task, and contextually appropriate emotion expression selection task, respectively.

hard). For "easy," we randomly selected sentences from the top 20 sentences with semantic similarity scores of 0.1 or lower. For "hard," to introduce more complexity than the easy level, we adjusted the semantic similarity score criterion to 0.4 and randomly selected sentences from the top 20 sentences with the highest scores.

5 Experiments and Results

We measured the performance using three different prompts, considering their influence. The detailed prompts can be found in Appendix D.

5.1 Baselines

The experiments were conducted using various open-source VLLMs. Specifically, factors such as modality alignment, model size, and LLMs were considered in the model selection. Modality alignment is a technique for effectively integrating and processing various types of data, such as text and images, in VLLMs. We analyzed key modality alignment techniques, including Direct Mapping (Liu et al., 2023), Q-Former (Li et al., 2023d), and Customization Perceiver (Alayrac et al., 2022; Awadalla et al., 2023). In addition, following gen-

erally known scaling laws, we thoroughly investigated how the emotion prediction performance of VLLMs interacted with other factors. To this end, we also conducted experiments on models with the same architecture but different sizes. The selected VLLMs included LLaVA-1.5 (Liu et al., 2023), MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2024), Qwen-VL-Chat (Bai et al., 2023), LLaVA-Next (Liu et al., 2024), and Otter (Li et al., 2023a). We selected various LLMs and model sizes and performed experiments on 11 VLLMs. Some high-performance models, such as GPT-4V, were excluded from the detailed analysis because their internal workings and parameter configurations have not been disclosed.

5.2 Main Results

Table 1 presents the performance based on the average values of the easy and hard difficulty levels. The individual performances for easy and hard can be found in Appendix B. We provide answers to the following questions according to the main experimental results:

Q1: What is the most influential factor in the emotion prediction performance of the model? Answer: **LLM.** Our experiments show that the most important factor in the emotion prediction performance of a model is the LLM itself. In particular, we observed that as the size of the LLM increased, the performance consistently improved across all models used in the experiments (InstructBLIP, LLaVA-1.5, LLaVA-Next, etc.). This trend was evident across all prompt types, including original, personality traits, speaking styles, and chain of thought (CoT). These results suggest that the LLM Backbone plays a more crucial role in predicting human emotions than focusing on specific image regions does. This aligns with existing research, indicating that the architecture and scaling of LLMs enhance the performance.

Q2: What is the most outstanding model architecture for emotion prediction? Answer: **InstructBLIP(FLAN 11B).** InstructBLIP(FLAN 11B) consistently achieved the highest performance in most cases. To verify whether these results were simply owing to the instruction-tuning dataset, we conducted a comparative experiment with InstructBLIP(Vicuna 13B), which was trained using the same data. Consequently, FLAN exhibited superior performance over Vicuna, indicating that the architecture of FLAN itself, rather than merely the tuning data, provides excellent emotion prediction performance.

Q3: Do additional persona information and CoT affect the emotion prediction performance of the model? Answer: **Yes.** The integration of persona information and CoT prompts influenced the emotion prediction performance of the model. The experimental results indicated that the effects of these elements varied depending on the model. Some models (e.g., LLaVA-1.5 and Qwen-VL-Chat) exhibited slight performance improvements when persona information or CoT prompts were added, whereas other models (e.g., InstructBLIP and LLaVA-Next) showed no significant differences or performance degradation. This suggests that persona information and CoT prompts may have different effects depending on the model architecture or pre-training data. However, considering that the overall performance improvement was not substantial, the effects of these elements appear to be limited. Therefore, future research should explore means of using persona information and CoT prompts more effectively.

6 Analysis

6.1 How do different prompts affect the overall emotion prediction performance?

We analyzed the performance for each emotion in the emotion prediction. As shown in Figure 2, all prompt types showed the highest performance in predicting the "joy" emotion, with the speaking styles prompt achieving the best result of 50.82%. This suggests that the tone and style of conversation play an important role in predicting positive emotions. The personality traits prompt also showed high performance in predicting "joy," at 49.98%, indicating that individual personality traits are crucial elements in understanding and expressing joy. These results demonstrate that the model can predict positive emotions more accurately based on the personality and speaking style of the speaker.

In contrast, all prompt types showed relatively lower performance in predicting "fear" than other emotions. In the case of the speaking styles prompt, the performance for predicting the "fear" emotion was the lowest among all emotions, at 39.58%, and similar trends were observed for the other prompt types. This indicates that predicting negative emotions such as fear is challenging. Fear may require complex and subtle contexts and the limitations of the model may be exposed when accurately predicting such emotions.

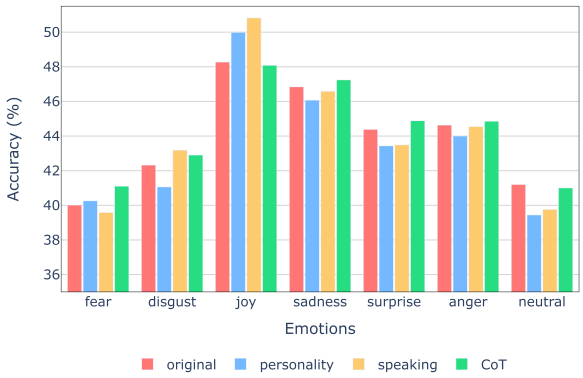

Figure 2: Comparison of emotion prediction performance across different prompt types (original, personality traits, speaking styles, and CoT).

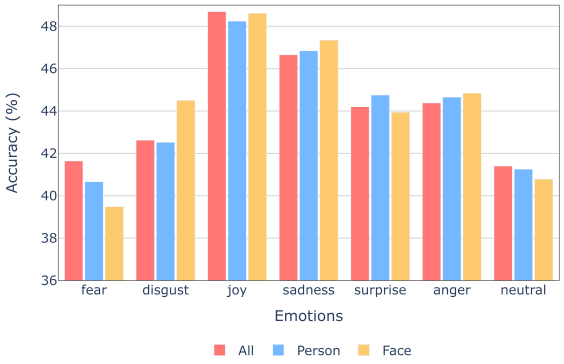

Figure 3: Changes in emotion prediction performance based on image scope (all, person, and face) for each emotion category.

In addition, the prediction of the "neutral" emotion showed relatively low performance across all prompt types, particularly in the personality traits prompt, which had the lowest performance at 39.44%. This suggests that individual personality traits may add complexity to the process of discerning emotional neutrality. Neutral emotions are difficult to predict owing to the absence of clear positive or negative signals, indicating that additional research is required to improve the model performance to respond in situations in which clear emotional signals are lacking.

6.2 How does emotion prediction performance differ based on the image scope?

We demonstrated the differences in emotion prediction performance based on the image scope, as shown in Figure 3. For "joy," high performance was observed across all scopes, with the "all" scope achieving the best result at 48.69%. However, the performance of the "face" and "person" scopes was not significantly different, at 48.61% and 48.24%, respectively. This suggests that various cues, such as facial expressions, individuals, and the overall context, may be equally important when predicting joy.

For "sadness," the "face" scope showed the highest performance at 47.34%, suggesting that facial expressions are a crucial factor in predicting sadness. However, for "fear," the "all" scope exhibited the highest performance, at 41.64%. This implies that the overall image information can be helpful in the prediction of fear because it is an emotion that arises in complex contents.

For "disgust," the "face" scope achieved the highest performance at 44.50%, whereas for "surprise," the "person" scope showed the highest performance at 44.75%. This indicates that facial expressions and posture or movements of an individual can play important roles in predicting disgust and surprise, respectively. For "anger," the performance difference between the image scopes was not significant, with the "face" scope showing a slightly higher level at 44.84%.

In contrast, "neutral" showed relatively low performance across all scopes, particularly in the "face" scope, which had the lowest performance at 40.78%. This suggests that facial expressions alone may not provide sufficient cues for predicting neutral emotions. The "all" scope showed the highest performance at 41.40%, but this low level suggests that more sophisticated context analysis may be necessary to predict neutral emotions accurately.

6.3 Is the emotion prediction performance of the model influenced by gender?

In this section, we analyze whether differences in emotion prediction performance occurred based on gender. The experimental results presented in Figure 4 clearly show how the emotion prediction performance varied depending on the gender of the subject that the model aimed to predict. The results revealed that the emotion prediction performance of female was higher than that of male for most emotions. Notably, for the "disgust" emotion, the prediction performance for females (54.21%) was significantly superior to that for males (34.78%). A detailed analysis is provided in Appendix E. For the "joy" and "surprise" emotions, the prediction performance for females was also higher at 50.59% and 46.19%, respectively, compared to males (46.63% and 41.68%, respectively). This

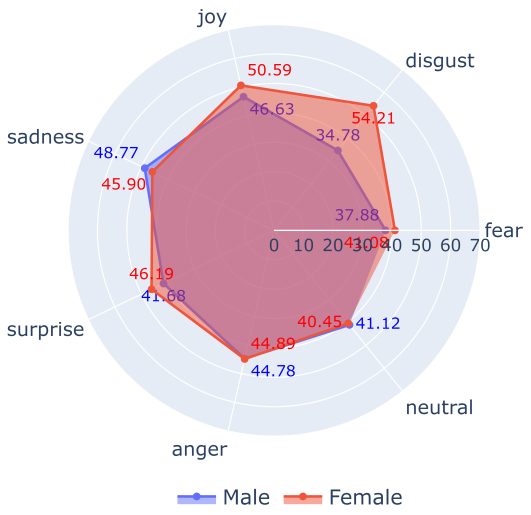

Figure 4: This radar chart illustrates the differences in emotion prediction performance based on the target gender.

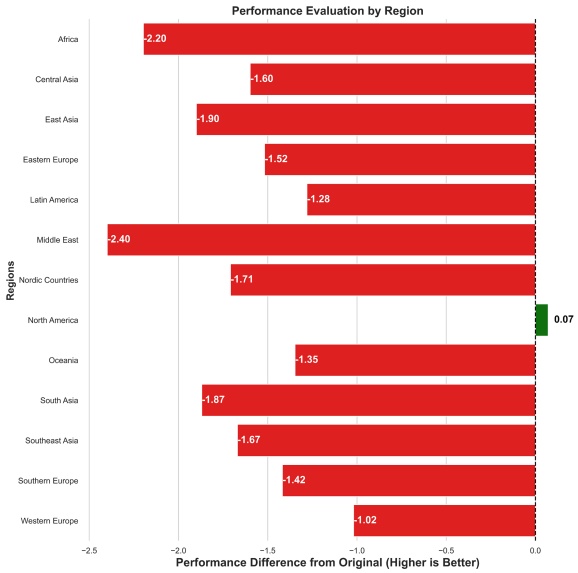

Figure 5: Changes in emotion prediction performance based on region, calculated according to the difference from the emotion prediction performance of the original prompt.

implies that positive emotions such as "joy" and "surprise" may be more prominently expressed by females.

In contrast, when recognizing the "sadness" emotion, the performance for males (48.77%) was higher than that for females (45.90%). This suggests that male emotional expressions may be better recognized by the model when identifying "sadness." In the recognition of the "anger" and "neutral" emotions, the performance difference between males and females was not significant, indicating that the expression differences of "anger" and "neutral" based on gender may be relatively small.

6.4 Do regional biases influence the emotion prediction performance?

The experimental results presented in Figure 5 clearly demonstrate the impact of regional biases on the emotion prediction performance of the model. According to the analysis, most regions showed a tendency for the model performance to degrade when regional persona information was added. In particular, for the Middle East and Africa, the performance decreased by -2.40% and -2.20%, respectively, compared to the original prompt, indicating that regional biases had a negative impact on the emotion prediction performance. Performance degradation was also observed for East Asia (-1.90%), South Asia (-1.87%), and Nordic countries (-1.71%).

In contrast, North America was the only region for which the performance improved by +0.07%. This suggests that the data used to train the model reflect the characteristics of the North American region relatively well. For Latin America and Western Europe, the performance decreases were relatively small, at -1.28% and -1.02%, respectively; however, they still appeared to be influenced by regional biases. Additional details are provided in Appendix F.

7 Conclusion

This study has systematically analyzed the key factors influencing the emotion prediction performance of VLLMs. The experimental results showed that the model architecture and size, particularly the LLM Backbone, had the most significant impact. The integration of persona information and CoT prompts exhibited varying effects depending on the model, and differences in the prediction performance were observed based on the image scope for each emotion. However, biases in emotion prediction performance based on gender and region were identified, indicating the need for efforts to mitigate these biases. Future research should focus on developing emotionally intelligent VLLMs by minimizing data and model biases using advanced dataset composition and model training methods.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00343989, Enhancing the Ethics of Data Characteristics and Generation AI Models for Social and Ethical Learning) and (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI).

Limitations

Although this study provides valuable insights into the factors that influence the emotion prediction performance of VLLMs, it has some limitations that should be acknowledged. First, we excluded high-performing models, such as GPT-4V, from our detailed analysis because their internal structures and model sizes have not been publicly disclosed. Although these models are likely to employ advanced architectures that can further our understanding of emotion prediction, their lack of transparency makes it difficult to analyze the specific factors that contribute to their performance systematically. However, as more information on these models becomes available, future research should investigate their emotion prediction capabilities in relation to the factors identified in this study.

Second, although our experiments revealed the presence of gender and regional biases in the emotion predictions of VLLMs, proposing comprehensive solutions to these biases is beyond the scope of this study. Addressing these biases is crucial for developing fair and unbiased VLLMs, and we strongly encourage future research to focus on mitigating these issues.

Finally, it is important to note that although the MELD dataset, which is based on the TV series *Friends*, reflects many real-world emotional situations, it may not capture the full range of emotions and contexts that are present in human interactions. Although TV shows are designed to mirror real life, they are ultimately scripted and may not always represent the spontaneity and complexity of real-world conversations. Future research could expand the scope of this study by incorporating datasets from diverse sources, such as real-world conversations, to validate and generalize our findings further.

Despite these limitations, we believe that our study provides a solid foundation for understanding the factors that influence the emotion prediction performance of VLLMs. We have identified key areas for future research and development in this field by systematically analyzing the effects of the model architecture, persona information, and various biases. As VLLMs continue to advance, it will be crucial to address these limitations and build emotionally intelligent models that can understand and respond to human emotions in a fair and unbiased manner.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Sander L Koole. 2010. The psychology of emotion regulation: An integrative review. *Cognition and emotion*, pages 138–177.

Richard S Lazarus. 2000. How emotions influence performance in competitive sports. *The sport psychologist*, 14(3):229–252.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023b. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023c. The good, the bad, and why: Unveiling emotions in generative ai. *arXiv preprint arXiv:2312.11111*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P Perry. 2002. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist*, 37(2):91–105.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Peter Salovey, John D Mayer, David Caruso, and Seung Hee Yoo. 2009. The positive psychology of emotional intelligence.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Dataset Statistics

We maintained the emotion ratios used in the original MELD dataset and sampled emotions to construct a final dataset of 1,112 instances. The proportions of sentiments and emotions used in the data composition can be found in Figure 6 and Figure 7.

B Main Results for Easy and Hard Difficulty Levels

In this section, we present the performance variations of the models based on the difficulty levels of the emotion prediction tasks. Table 2 showcases the detailed results for both the easy and hard difficulty levels. These findings provide valuable insights into the capabilities of the models in understanding and processing emotions in diverse situational complexities.

C Model Hyperparameter Configuration

We conducted experiments on all models using the same hyperparameter values. Specifically, we set the do_sample parameter uniformly to False across all experiments.

D Prompt details

We conducted tests using the following three prompts to reduce the influence of prompts on the models:

- *Given a conversation involving multiple speakers and an associated image, select the most appropriate statement for the final speaker in the conversation. Consider the context, sentiment, and emotions conveyed in the dialogue*

and the image to identify the correct answer. Only one of the options is correct, and the others are incorrect.

- *Using a given dialogue involving multiple speakers and a related image, identify the most suitable reply for the last speaker based on the overall Tone, details in the conversation, and visual elements from the image. Remember, only one response is correct; the others do not fit the context as well.*

- *Based on the interaction among multiple speakers and the visual cues from the accompanying image, deduce which statement would most accurately reflect the final speaker's intended communication. Assess the tone, implied sentiments, and emotional context presented both verbally and visually. Only one option is the right answer; all others are incorrect.*

In addition, we used corresponding persona information as an additional input to analyze the differences in emotional prediction performance according to personality traits and speaking styles. We utilized the Friends Fandom Wiki to generate persona information, providing the relevant data as input to GPT-4 to create the persona information. Considering the maximum token limit for specific models, we only added two persona information inputs. The persona information input for personality traits follows this format:

Last speaker's personality traits:
1. [Personality trait]
2. [Personality trait]

Similarly, the speaking styles are input as follows:

Last speaker's speaking styles:
1. [Speaking style]
2. [Speaking style]

The overall prompt can be found in Figure 9.

E Analysis of "Disgust" Emotion Prediction Differences by Gender

The conversational samples in Figures 10 and 11 reveal notable differences in how males and females express the emotion of "disgust." In the female samples, disgust is often expressed through strong exclamations such as "Oh my God!" and "Ewww!" (Figure 10, Samples 1 and 3). These expressions suggest a more overt and emphatic display of the

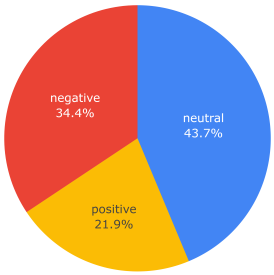

Figure 6: Sentiment distribution used for sentiment analysis

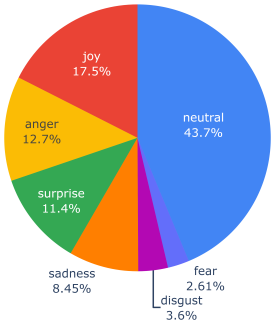

Figure 7: Emotion distribution used for emotion analysis

"disgust" emotion by females. In addition, female characters tend to provide more detailed descriptions of the disgusting situation, such as "She's got her tongue in his ear" (Figure 10, Sample 2), which vividly conveys their sense of revulsion.

In contrast, the male samples show a relatively more subdued expression of "disgust." For instance, in Figure 11, Sample 1, the male character expresses his aversion to drinking breast milk in a more matter-of-fact manner, stating, "Not even if Carol's breast had a picture of a missing child on it." While still conveying disgust, the expression is less emotionally charged compared to the female samples. Similarly, in Figure 11, Sample 3, the comment by the male character, "OK, is there a mute button on this woman?" suggests annoyance and disgust, but lacks the same level of overt expression that appears in the female samples.

These differences in the expression of disgust between males and females could potentially explain the higher performance in predicting "disgust" for female (54.21%) compared to male (34.78%). The

more explicit and emphatic expressions of disgust by females may provide clearer cues for the VLLMs to identify the emotion accurately.

However, it is important to acknowledge the limitations of this analysis. The conversational samples provided, while based on a TV show reflecting many real-world situations, do not fully capture the entire spectrum of "disgust" emotion expressions that occur in real-life interactions. In addition, the differences observed in these specific samples may be influenced by individual character traits and situational contexts, rather than being solely attributable to gender.

A more comprehensive study with a larger and more diverse dataset would be necessary to draw more definitive conclusions regarding gender-based differences in "disgust" emotion expression. Such a study should consider various factors, including individual personality traits, cultural backgrounds, and conversational contexts, to determine whether the observed differences are truly representative of gender-based patterns or whether other factors play a more significant role.

In summary, while the analysis of the provided conversational samples suggests potential differences in how males and females express disgust, further research is required to establish the extent to which these differences are generalizable across a wider population and to determine the relative influence of gender compared to other factors in shaping "disgust" emotion expression.

F Regional Bias Problem in Emotion Prediction of VLLMs

This study identified a general trend towards decreased emotion prediction performance when persona prompts containing regional information were provided to the models. This suggests that the models may inherently hold prejudices or stereotypes towards specific regions.

The prompts used in the experiments were structured as follows:

Last speaker's characteristics:

1. The last speaker has lived in ## throughout their life, deeply rooted in the language, religion, and customs of that region.

2. The last speaker uses the communication style commonly employed in ## to interact with others.

In the above, *##* was replaced with the corresponding region name. These prompts provided the model with the information that the last speaker is from a specific region and is deeply connected to the language, religion, customs, and communication style of that region.

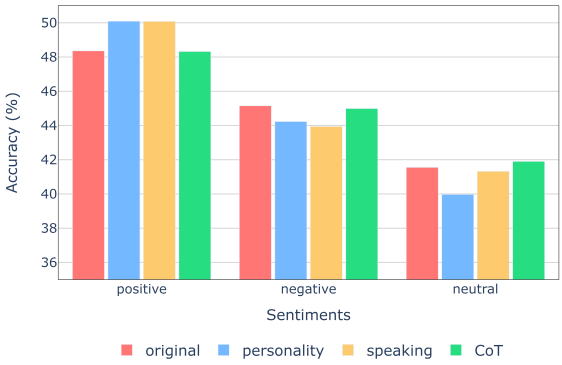

Figure 8: Changes in sentiment prediction perofrmance according to various prompts

However, the research results showed that the emotion prediction performance of the model deteriorated when such regional information was provided, demonstrating the possibility that the models harbor stereotypes or biases towards specific regions. These models may make inappropriate assumptions based on the regional information provided through prompts, leading to inaccurate emotion predictions.

This finding is directly related to the fairness and bias issues in VLLMs. If the models make biased predictions about specific regions, this can lead to unfair treatment of individuals in those regions. Therefore, future research is necessary to minimize such biases and enhance the fairness of the models.

G How do different prompts affect the overall sentiment prediction performance?

We analyzed the impact of various prompt types on the overall sentiment prediction performance of the models systematically, based on the experimental results presented in Figure 8. The results revealed that the inclusion of persona information, such as personality traits and speaking styles, influenced the sentiment prediction performance significantly, with notable variations observed across different sentiments. For the recognition of "positive" sentiments, the models exhibited a substantial improvement in performance when persona information was incorporated. In contrast, for the recognition of "negative" sentiments, the original prompt, which did not include any persona information, recorded the highest performance. When it comes to "neu-

tral" sentiments, the CoT prompt, which involved a systematic thought process, demonstrated the highest performance.

The experimental results showed that the inclusion of persona information had mixed effects on the sentiment prediction performance of the models, with the overall performance improvement being limited. While the personality prompt showed promising results for positive sentiments, it did not benefit all sentiment prediction tasks consistently. Similarly, the speaking styles prompt, although effective for positive sentiments, did not yield significant improvements in the recognition of negative or neutral sentiments. These findings suggest that the impact of persona information on sentiment prediction performance varies depending on the specific emotion being analyzed.

Our analysis highlights the importance of considering the interplay between persona information and sentiment prediction in conversational contexts. While the inclusion of personality traits and speaking styles can enhance the models' understanding of certain sentiments, such as positive sentiments, its impact is not uniform across all sentiment categories. Further research is needed to explore more sophisticated approaches for integrating persona information into sentiment prediction tasks, taking into account the nuances and challenges associated with different emotional expressions.

Instruction
Given a conversation involving multiple speakers and an associated image, select the most appropriate statement for the final speaker in the conversation. Consider the context, sentiment, and emotions conveyed in the dialogue and the image to identify the correct answer. Only one of the options is correct, and the others are incorrect.

Historical content
Conversations:
Speaker 1: "Hey Estelle, listen"
Speaker 2: "Well! Well! Well! Joey Tribbiani! So you came back huh? They"
Speaker 1: "What are you talking about? I never left you! You've always been my agent!"
…
Speaker 1: “Yeah!“
Speaker 2:

Personality trait
Last speaker's personality traits:
1. She communicates directly and openly, expressing her opinions without hesitation.
…

Speaking style
Last speaker’s speaking styles:
1. She communicates impressively with a strong voice and forceful tone.
…

Options
Last speaker's potential responses:
(A) It's, I mean, it's nothing, I'm fine.
(B) Oh... that's too bad. Bye bye.
(C) Oh, I'm so sorry.
(D) Oh well, no harm, no foul.

CoT
Let's think step by step.

Answer
The last speaker's response is most likely to be

Figure 9: Example of the prompt template used for testing. This figure illustrates the detailed structure of the prompt used in our experiments, including sections for instruction, historical content, personality traits, speaking styles, response options, and the CoT. This comprehensive prompt format ensures that the model evaluates multiple aspects of context and persona information to determine the most appropriate response.

Model	LLM	Easy Tone			Easy Emotion			Easy Context			Hard Tone			Hard Emotion			Hard Context			Avg.
		All	Person	Face	All	Person	Face	All	Person	Face	All	Person	Face	All	Person	Face	All	Person	Face	
Prompt type										*Original*										
InstructBLIP	Vicuna (13B)	46.07	46.52	44.93	44.99	45.17	44.06	46.76	47.00	46.67	34.38	33.78	33.27	35.58	36.48	36.84	34.68	35.19	35.64	40.45
LLaVA-1.5	Vicuna (13B)	58.87	58.63	57.61	55.34	55.19	55.13	55.37	55.64	54.71	41.91	41.67	41.94	42.63	42.69	42.36	42.00	41.13	40.29	49.06
LLaVA-Next	Vicuna (13B)	59.23	57.76	58.27	56.24	55.07	54.77	56.50	56.41	55.19	44.78	45.08	44.30	43.50	43.29	42.36	43.44	43.11	43.08	50.13
InstructBLIP	FLAN (11B)	63.16	63.34	63.37	64.09	64.48	64.66	62.44	62.53	63.28	49.04	49.37	49.82	49.40	49.01	49.19	48.26	48.14	48.17	56.20
LLaVA-1.5	Vicuna (7B)	43.76	43.94	43.14	41.55	42.00	41.88	41.34	40.68	40.53	32.46	32.88	31.77	31.44	31.26	31.47	30.10	30.28	29.11	36.65
LLaVA-Next	Vicuna (7B)	53.27	53.39	52.70	52.64	52.85	52.70	51.92	51.89	51.50	39.54	38.73	39.51	39.00	38.67	38.31	38.40	38.07	37.56	45.59
LLaVA-Next	Mistral (7B)	54.29	54.32	53.30	51.56	52.37	51.38	52.22	51.53	51.47	41.43	40.83	41.67	41.73	40.56	41.67	40.77	40.53	40.11	46.76
Qwen-VL-Chat	Qwen (7B)	44.06	44.57	44.84	42.57	42.54	42.36	42.39	42.36	42.51	35.49	34.41	35.49	35.19	35.31	34.68	33.12	33.57	32.61	38.78
MiniGPT-4	Vicuna (7B)	28.99	29.68	29.80	28.63	28.33	28.09	26.80	27.46	27.52	26.17	26.11	27.34	27.01	27.37	26.35	26.11	26.08	26.62	27.47
Otter	MPT (7B)	40.20	39.42	39.54	40.89	40.14	40.56	39.69	39.78	40.11	35.79	35.88	36.30	36.27	36.60	37.23	34.53	34.68	33.90	37.86
InstructBLIP	FLAN (3B)	60.25	60.07	60.31	59.14	58.78	59.35	60.40	60.37	60.07	43.56	43.76	42.42	44.57	44.36	43.68	41.10	40.92	40.05	51.28
Prompt type										*Personality traits*										
InstructBLIP	Vicuna (13B)	44.87	44.75	43.62	44.72	44.30	43.65	44.99	44.78	43.91	35.04	34.98	34.38	36.09	36.24	35.49	34.62	34.41	34.38	39.73
LLaVA-1.5	Vicuna (13B)	59.02	58.75	57.67	55.04	55.61	54.62	55.46	55.76	54.14	43.17	43.26	43.32	43.62	43.59	43.91	43.41	43.74	43.05	49.84
LLaVA-Next	Vicuna (13B)	55.79	54.68	54.62	51.68	51.14	50.75	54.47	54.41	52.85	44.09	43.41	42.48	42.78	42.33	41.91	42.63	42.54	41.70	48.02
InstructBLIP	FLAN (11B)	60.88	61.36	60.31	62.62	63.31	62.53	60.46	60.70	60.61	48.86	48.65	48.08	47.27	46.58	46.55	45.65	45.41	46.28	54.23
LLaVA-1.5	Vicuna (7B)	43.82	43.82	43.56	40.44	40.59	41.40	41.28	40.68	40.98	31.12	31.59	31.24	31.12	31.41	31.26	29.80	29.80	29.80	36.31
LLaVA-Next	Vicuna (7B)	51.38	51.80	51.11	51.53	51.32	51.50	52.46	52.52	51.68	38.58	38.04	38.28	38.25	37.53	38.16	37.38	37.44	36.27	44.73
LLaVA-Next	Mistral (7B)	51.83	51.83	51.83	50.54	51.20	50.15	50.33	50.30	50.36	40.71	40.32	40.62	40.53	39.81	40.29	39.93	40.35	39.90	45.60
Qwen-VL-Chat	Qwen (7B)	44.39	43.94	44.15	42.51	42.78	42.84	41.34	41.04	42.00	35.52	36.24	35.91	34.35	34.38	33.93	33.93	34.26	34.08	38.75
MiniGPT-4	Vicuna (7B)	28.99	29.80	30.40	30.22	30.49	30.64	28.87	29.62	28.93	27.31	28.39	28.24	26.56	26.53	27.46	28.45	28.60	28.39	28.77
Otter	MPT (7B)	41.28	41.28	40.89	41.67	41.07	41.52	41.82	41.40	41.94	36.24	36.27	36.72	37.02	37.56	37.38	33.93	34.35	34.71	38.73
InstructBLIP	FLAN (3B)	58.12	58.18	57.82	56.29	56.09	56.50	58.33	58.03	58.06	41.37	41.25	40.86	41.94	41.94	40.68	38.88	38.52	38.37	48.95
Prompt type										*Speaking styles*										
InstructBLIP	Vicuna (13B)	44.69	44.90	43.76	45.20	45.20	43.94	44.90	45.05	44.54	35.67	35.34	35.16	35.46	36.18	35.25	34.35	34.14	34.44	39.90
LLaVA-1.5	Vicuna (13B)	57.94	57.70	56.29	54.44	55.04	53.96	55.22	55.01	53.93	43.02	42.78	42.39	43.91	43.94	43.62	42.33	42.81	41.76	49.23
LLaVA-Next	Vicuna (13B)	55.19	54.35	54.59	51.44	51.53	50.21	53.12	52.46	52.43	45.71	44.96	44.75	42.30	42.36	42.24	41.82	41.55	41.82	47.94
InstructBLIP	FLAN (11B)	62.35	63.43	62.65	63.52	64.12	64.27	62.20	62.56	62.95	49.07	49.58	49.61	47.24	47.69	47.63	47.00	46.37	46.91	55.51
LLaVA-1.5	Vicuna (7B)	44.54	44.30	44.03	43.38	43.08	42.99	41.58	41.82	41.37	33.03	32.85	32.13	32.76	31.83	32.43	31.00	30.91	30.70	37.48
LLaVA-Next	Vicuna (7B)	52.07	52.31	51.26	50.60	51.08	51.17	51.98	52.31	51.80	38.19	38.64	38.91	38.52	38.22	38.76	37.65	37.02	37.38	44.88
LLaVA-Next	Mistral (7B)	53.93	54.17	52.76	51.62	52.22	50.90	50.72	51.38	50.66	41.82	41.40	41.64	41.52	40.71	41.16	41.52	41.16	41.16	46.69
Qwen-VL-Chat	Qwen (7B)	43.97	44.36	44.90	43.71	43.88	43.65	42.12	42.54	42.54	35.94	36.48	35.76	34.38	34.20	33.78	34.74	34.62	34.53	39.23
MiniGPT-4	Vicuna (7B)	30.76	30.58	30.16	31.29	31.65	31.86	29.86	30.25	30.28	27.49	28.03	28.06	27.46	27.85	28.24	28.93	29.05	28.90	29.48
Otter	MPT (7B)	42.03	41.43	42.21	41.19	40.41	41.37	41.34	41.73	41.55	36.66	36.51	36.69	37.68	37.53	37.80	34.86	34.92	35.82	38.98
InstructBLIP	FLAN (3B)	57.73	58.06	57.52	56.56	56.74	57.04	57.64	58.33	57.76	41.22	41.64	41.04	42.06	41.64	40.53	38.52	38.58	38.10	48.93
Prompt type										*CoT*										
InstructBLIP	Vicuna (13B)	46.13	46.46	44.66	45.74	45.53	44.09	46.79	46.82	45.89	34.92	34.47	33.69	37.29	36.51	36.39	35.55	35.55	34.92	40.63
LLaVA-1.5	Vicuna (13B)	57.52	57.37	56.77	54.95	54.98	54.80	55.25	54.95	54.35	41.13	40.71	41.10	43.35	43.29	43.32	41.01	40.92	40.08	48.66
LLaVA-Next	Vicuna (13B)	57.22	55.85	55.49	53.21	51.98	51.86	56.18	56.18	54.86	45.20	44.96	43.91	43.50	43.44	42.39	43.56	43.02	42.09	49.16
InstructBLIP	FLAN (11B)	62.83	63.07	62.80	63.19	63.70	63.97	62.23	62.53	62.41	49.04	49.16	49.25	48.50	48.17	48.38	47.63	47.60	47.69	55.67
LLaVA-1.5	Vicuna (7B)	47.84	47.93	46.79	47.00	46.94	46.94	45.77	45.35	44.51	35.91	36.39	36.21	35.88	35.73	35.22	33.48	34.02	32.28	40.78
LLaVA-Next	Vicuna (7B)	53.45	53.21	51.98	52.46	52.58	51.47	51.92	52.46	51.38	38.97	38.88	39.12	38.76	38.07	37.62	36.63	37.05	36.57	45.14
LLaVA-Next	Mistral (7B)	55.22	54.47	52.91	52.46	53.57	51.95	53.66	52.40	52.16	41.79	41.31	41.76	42.03	41.10	41.13	41.67	40.71	38.73	47.17
Qwen-VL-Chat	Qwen (7B)	43.47	43.56	43.79	41.37	40.47	41.43	41.10	40.62	40.11	35.61	34.92	35.10	33.63	34.23	34.68	32.61	32.67	32.52	37.89
MiniGPT-4	Vicuna (7B)	29.62	29.68	29.44	28.09	27.61	27.46	27.22	27.25	26.86	26.17	26.32	26.44	26.56	27.01	26.20	25.57	25.09	24.97	27.09
Otter	MPT (7B)	40.05	38.91	38.79	39.00	38.76	38.70	38.88	39.30	38.94	35.67	35.40	35.46	36.12	35.61	36.06	33.69	34.02	33.45	37.05
InstructBLIP	FLAN (3B)	60.73	60.34	60.67	59.71	59.41	60.01	60.28	60.49	60.25	43.56	43.68	43.23	44.87	44.69	43.79	41.61	41.37	40.71	51.64

Table 2: Comparative performance analysis of VLLM models using different prompt types (original, personality traits, speaking styles, and CoT) for both easy and hard difficulty levels. The results, presented as accuracy scores, are averaged across three distinct prompts for each prompt type and are reported separately for Easy and Hard difficulties, allowing for a more detailed comparison of model performance across different complexity levels.

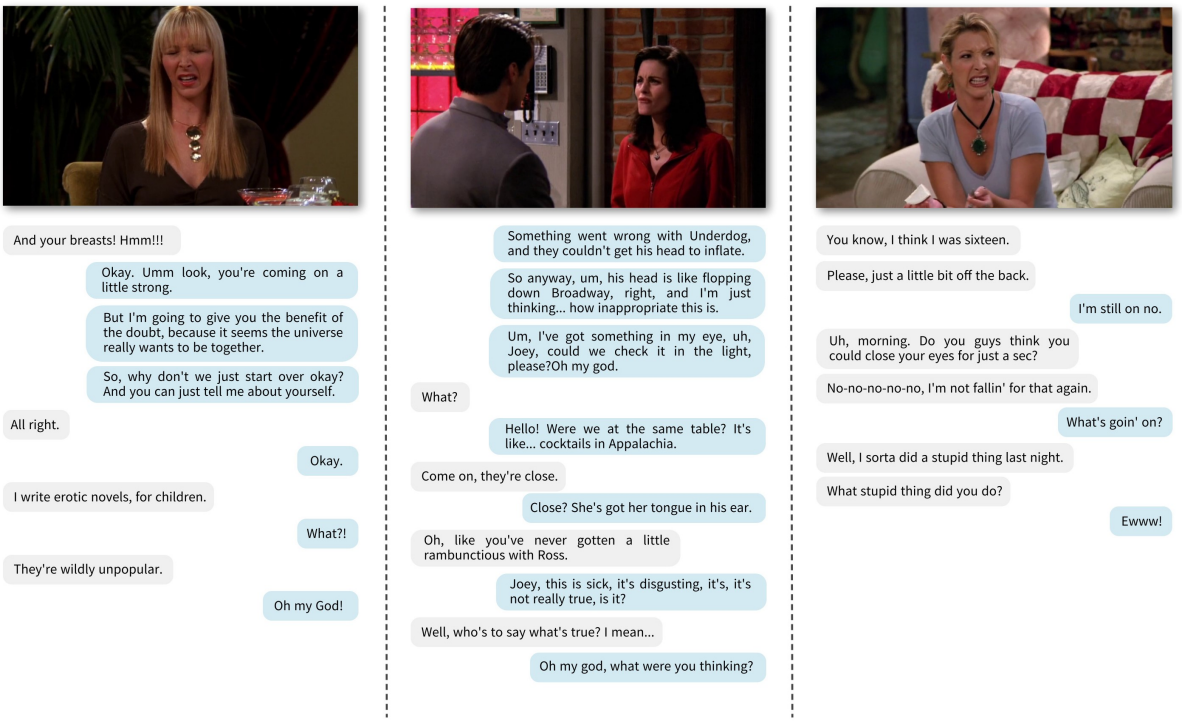

Figure 10: Example dialogues in which female characters express "disgust" in the final utterance.

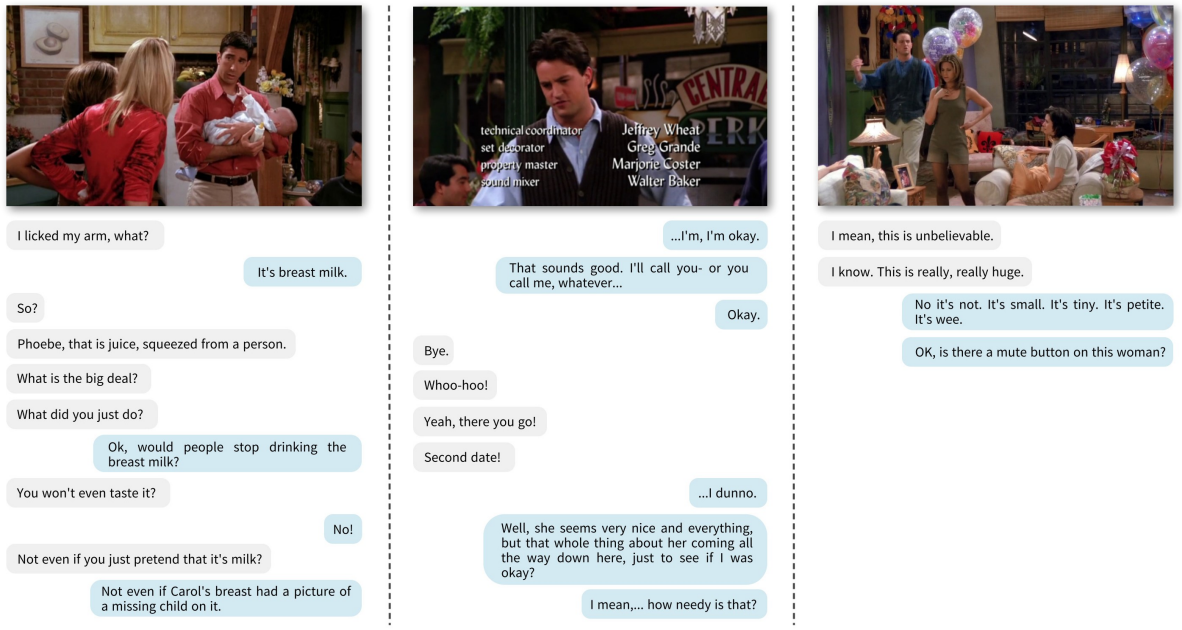

Figure 11: Example dialogues in which male characters express "disgust" in the final utterance.