# Bias in Emotion Recognition with ChatGPT

Naoki Wake<sup>1</sup>, Atsushi Kanehira<sup>1</sup>, Kazuhiro Sasabuchi<sup>1</sup>, Jun Takamatsu<sup>1</sup>, and Katsushi Ikeuchi<sup>1</sup>

Abstract—This technical report explores the ability of Chat-GPT in recognizing emotions from text, which can be the basis of various applications like interactive chatbots, data annotation, and mental health analysis. While prior research has shown ChatGPT's basic ability in sentiment analysis, its performance in more nuanced emotion recognition is not yet explored. Here, we conducted experiments to evaluate its performance of emotion recognition across different datasets and emotion labels. Our findings indicate a reasonable level of reproducibility in its performance, with noticeable improvement through fine-tuning. However, the performance varies with different emotion labels and datasets, highlighting an inherent instability and possible bias. The choice of dataset and emotion labels significantly impacts ChatGPT's emotion recognition performance. This paper sheds light on the importance of dataset and label selection, and the potential of fine-tuning in enhancing ChatGPT's emotion recognition capabilities, providing a groundwork for better integration of emotion analysis in applications using ChatGPT.

#### I. Introduction

The ability to recognize emotions from text is a fundamental aspect of various applications ranging from interactive chatbots [1], [2], annotated data generation [3], text analysis [4], and mental health analysis [5]–[7]. As chatbots become increasingly sophisticated, understanding the emotional nuance within user inputs can significantly enhance the conversational experience by enabling more empathetic and contextually appropriate responses. Similarly, in the realm of mental health, analyzing textual data for emotional content can provide crucial insights into an individual's mental health conditions.

In recent years, the advancements in Large Language Model (LLM), notably models like ChatGPT, have shown promise in various Natural Language Processing (NLP) tasks, including emotion recognition. Early studies have demonstrated the capability of ChatGPT in basic sentiment analysis tasks where the primary goal is to distinguish between positive and negative sentiments [7]–[10]. Beyond this binary classification a few reports delved into a nuanced understanding of emotions like joy, sadness, anger, and surprise and showed that ChatGPT shows reasonable performance in such detailed emotion analysis [5], [11] in a zero-shot and few-shot prompting conditions.

However, one of the caveats associated with LLMs is their tendency to exhibit biases in the generated outputs. These biases often have roots in the non-uniformity of their training data and the labeling methodologies employed [12]. Consequently, it is likely that potential biases manifest in the nuances of emotional recognition. While understanding these specific tendencies in emotion recognition could offer insights for the responsible deployment of ChatGPT in real-world applications, detailed studies on this particular aspect remain limited.

Using ChatGPT as a representative LLM, this study delves into its emotion recognition capabilities across various datasets and emotion labels. In experiments that solely relied on prompting (i.e., a zero-shot condition), ChatGPT exhibited commendable emotion recognition performance, consistent with previous research findings. However, we observed disparities in performance across different emotion labels, and intriguingly, the same label sometimes yielded varied results across datasets. While fine-tuning generally enhanced the model's recognition abilities, inconsistencies among labels persisted. A notable observation was the sensitivity of the model to slight label name changes (e.g., from "happiness" to "happy"). Despite these fluctuations, ChatGPT's responses remained largely consistent, and the labeling context had minimal influence on its overall performance. Our findings not only elucidate ChatGPT's strengths and limitations in emotion recognition but also set the stage for future investigations focused on optimizing performance and minimizing biases in ChatGPT-driven emotion recognition tasks.

## II. METHODS

## A. ChatGPT and prompts

ChatGPT is an LLM developed by OpenAI [13]. One of its distinct features is its capability to generate text in a conversational manner. This interactive nature allows users to leverage the model for a wide range of NLP tasks using conversational texts, which is called a prompt. The methodology of crafting prompts to obtain desired outputs from the off-the-shelf model without additional dataset (i.e, zero-shot setting) is referred to as prompt engineering [10]. Additionally, there is a methodology that employs a dataset of expected input-output pairs to fine-tune the model's behavior. In this research, we employed the gpt-3.5-turbo, a fixed version of the GPT model that is publicly available for fine-tuning. This version represents the latest iteration of ChatGPT that publicly supports fine-tuning to date. We assessed its efficacy in both prompt engineering and fine-tuning conditions.

We evaluate ChatGPT's ability to discern emotions in complex scenarios based on the emotion recognition in conversation (ERC) task [14]. In ERC, the goal is to recognize the emotion of each utterance within a conversation from a predefined set of emotion categories. In this paper, we assume an ongoing conversation scenario where future conversation information is unobserved. We estimate the emotion label for the current utterance based solely on the history of past conversations. That is, given a conversation history  $C = \{u_1, u_2, \ldots, u_t\}$  consisting of a sequence of t utterances, the task is to recognize the emotion label to  $u_t$ , referred to as  $y_t$ . Here,  $y_t \in E$ , where E is the set of predefined emotion labels.

For the zero-shot setting, we designed a prompt based on a previous study [5]:

<sup>&</sup>lt;sup>1</sup>Applied Robotics Research, Microsoft, Redmond, WA 98052, USA naoki.wake@microsoft.com

TABLE I
THE DETAILED INFORMATION OF DATASETS FOR EMOTION RECOGNITION

Dataset	Data Source	Conv./Utter.	Emotion Label
IEMOCAP	Multimodal Act (only text modality)	31/1,622	neutral, sad, anger, happy, frustrated, excited
MELD	Friends TV Show	280/2,610	neutral, sadness, anger, disgust, fear, joy, surprise
EmoryNLP	Friends TV Show	85/1,328	neutral, sad, mad, scared, powerful, peaceful, joyful
DailyDialog	Human Written Scripts	1,000/7,740	neutral, happiness, surprise, sadness, anger, disgust, fear

You are an emotion analyzer capable of understanding the sentiment within a text. Previous dialogue (sequence of pairs of speaker and the utterance): "[Dialogue History]". (Your task) Consider this dialogue to assign one emotion label to the next utterance: "[Query]". Only one label. Only from this emotion list: [Label List].,

where [Dialogue History] denotes the history of dialogue containing a list of speaker names and utterances, [Query] denotes a target utterance with the speaker name, and [Label List] denotes emotion labels for a dataset. When using a dataset without speaker IDs, the name was denoted as 'unknown.'

#### B. Dataset and Evaluation Metrics

We evaluates the performance of emotion recognition by ChatGPT across five datasets: IEMOCAP [15], MELD [16], EmoryNLP [17], DailyDialog [18]. For the IEMOCAP multimodal dataset, we focus solely on the text modality. The detailed information of those dataset is summarized in Table I.

In our experiments, we adhered to the established metrics in line with the prevailing approach in the literature. Specifically, for the IEMOCAP, MELD, and EmoryNLP datasets, we employed the Weighted-F1 metric. For the DailyDialog dataset, considering the disproportionately high occurrence of neutral utterances, we utilized the micro-F1 metric, excluding the neutral category. It is noteworthy that our analysis was not limited to overall performance but also we also examined the recognition performance for each individual emotion label, to explore the specific tendencies in emotion recognition.

#### C. Fine-tuning

When fine-tuning ChatGPT, we used the prompts from the zero-shot experiments along with their corresponding correct labels. To test with a more generalized model, we combined four distinct datasets previously mentioned for fine-tuning, and tested the model across the four datasets. We utilized data from the validation sets of each dataset: 1365 samples from IEMOCAP, 1109 from MELD, 1344 from EmoryNLP, and 961 from DailyDialog. The distribution of the labels are shown in Fig.1

## III. RESULTS

#### A. Main Results

The overall performance in emotion recognition significantly exceeded chance levels, aligning with the findings from previous studies (TableII,III,IV,V). However, there was noticeable variance when examining performance based on

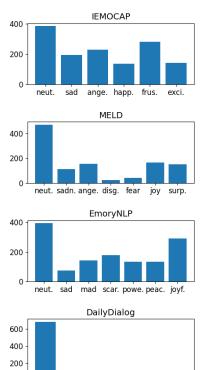


Fig. 1. The distribution of labels used for fine-tuning ChatGPT. For the DailyDialog dataset, we eliminated the neutral labels due to its occurrence.

sadn. ange.

happ. surp.

individual emotion labels. For example, in the zero-shot experiments on the EmoryNLP dataset, the recognition performance (F1 score) for 'joyful' was 51%, while it was below 5% for 'peaceful' and 'powerful.'

Fine-tuning generally observed to enhance performance (Fig.2). Interestingly, despite the EmoryNLP dataset using text based on the same TV series as MELD, the impact of fine-tuning wasn't as pronounced.

Despite the variability in the labels of the datasets used for fine-tuning, a strong correlation was observed between the performance in zero-shot and fine-tuning (correlation coefficient of 0.81). This result suggests that while fine-tuning improves performance, it still retains inherent tendencies.

To explore the potential semantic relationship between recognition performance and label tendencies, we classified the labels into three overarching categories: positive, neutral, and negative (Fig.3). However, our analysis did not reveal any discernible correlation between recognition performance and the semantic nature of the labels.

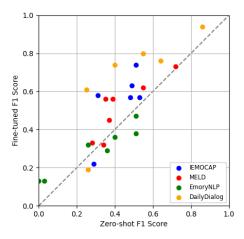


Fig. 2. Comparison of emotion recognition performance between zero-shot and fine-tuned models across four datasets.

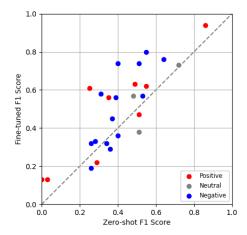


Fig. 3. Comparison of emotion recognition performance between zero-shot and fine-tuned models colored with positive/neutral/negative labels.

## B. Sensitivity to Label Text

From these results, we observed that recognition performance varied significantly across datasets, even for similar labels. For instance, in the DailyDialog dataset, the recognition performance for 'happiness' was around 94%, while in the IEMOCAP dataset, the recognition performance for 'happy' was only about 22% (see the fine-tuning columns in TableII and V). Notably, this tendency is observed in the zero-shot condition, as well. While these differences may be influenced by the unique characteristics of the sentences contained within each dataset, we hypothesized that there might be sensitivity to the label text. To test this, we changed the label name in the IEMOCAP dataset from 'happy' to 'happiness' and applied the same model to evaluate its performance.

As a result, the recognition performance for 'happiness' improved to 43% (TableVI). Additionally, the recognition performance for other labels did not suffer significantly, indicating that changing the label improved the recognition performance for the 'happy' label without compromising the

performance of other labels. This result suggests a sensitivity of the language model to the specific text of the labels.

## C. Reproducibility of the Performance

As an additional experiment, we investigated the reproducibility of ChatGPT's results using Cohen's Kappa score [19]. Under the zero-shot condition, when applying the ChatGPT model twice to the MELD test data, the Kappa score was 0.65. This suggests that we can expect reasonable reproducibility when applying the model multiple times.

## D. Dataset selection

In the experiment, the model was fine-tuned on a blend of all datasets. This dataset mixture can prevent over-fitting to a specific dataset, but it also has the potential to significantly affect the model performance. To gain insights on the impact of dataset mixture on the main result, we compared the results of fine-tuning on a single dataset against those trained on the mixed dataset, using the MELD dataset as an example. The model trained on a single dataset performed better for most labels, which is reasonable. However, the difference was at most about 7 percent, which is marginal compared to the difference across labels (49 percent at most), and there was no significant difference in the trend of performance differences between labels. These results suggest that the model's recognition bias is less likely to be affected by the scope of the dataset.

TABLE II
EMOTION RECOGNITION METRICS (IEMOCAP)

E di	zero-shot			fine-tuning			
Emotion	Precision	Recall	F1	Precision	Recall	F1	
neutral	0.36	0.70	0.48	0.67	0.49	0.57	
sad	0.79	0.38	0.51	0.85	0.65	0.74	
anger	0.70	0.20	0.31	0.47	0.74	0.58	
happy	0.33	0.25	0.29	0.24	0.20	0.22	
frustrated	0.52	0.54	0.53	0.53	0.62	0.57	
excited	0.65	0.39	0.49	0.62	0.64	0.63	
macro-average	0.56	0.41	0.43	0.56	0.56	0.55	
std	0.19	0.19	0.11	0.21	0.19	0.17	
weighted-f1 (Ours)	-	-	0.46	-	-	0.58	
weighted-f1 [11]	-	-	0.45	-	-	-	
weighted-f1 [5]	-	-	0.53	-	-	_	

Bold denotes higher values between zero-shot and fine-tuning.

TABLE III
EMOTION RECOGNITION METRICS (MELD)

	ze	ro-shot		fine-tuning		
Emotion	Precision	Recall	F1	Precision	Recall	F1
neutral	0.64	0.84	0.72	0.79	0.67	0.73
sadness	0.48	0.31	0.37	0.43	0.48	0.45
anger	0.65	0.28	0.39	0.53	0.59	0.56
disgust	0.33	0.34	0.34	0.37	0.28	0.32
fear	0.22	0.38	0.28	0.33	0.33	0.33
joy	0.55	0.55	0.55	0.61	0.62	0.62
surprise	0.58	0.25	0.35	0.47	0.69	0.56
macro-average	0.49	0.42	0.43	0.50	0.52	0.51
std	0.16	0.21	0.15	0.16	0.17	0.15
weighted-f1 (Ours)	_	_	0.57	_	-	0.63
weighted-f1 [11]	-	-	0.57	_	-	-
weighted-f1 [5]	-	-	0.61	_	-	-

TABLE IV
EMOTION RECOGNITION METRICS (EMORYNLP)

Emotion	zero-shot			fine-tuning		
EIIIOUOII	Precision	Recall	F1	Precision	Recall	F1
neutral	0.37	0.81	0.51	0.42	0.35	0.38
sad	0.41	0.32	0.36	0.33	0.27	0.29
mad	0.42	0.38	0.40	0.34	0.38	0.36
scared	0.38	0.19	0.26	0.29	0.36	0.32
powerful	0.00	0.00	0.00	0.16	0.11	0.13
peaceful	0.22	0.02	0.03	0.17	0.10	0.13
joyful	0.54	0.48	0.51	0.39	0.57	0.47
macro-average	0.34	0.31	0.29	0.30	0.31	0.30
std	0.17	0.28	0.21	0.10	0.16	0.13
weighted-f1 (Ours)	-	_	0.35	_	_	0.33
weighted-f1 [11]	_	-	0.37	-	-	_
weighted-f1 [5]	-	-	0.33	-	-	_

TABLE V
EMOTION RECOGNITION METRICS (DAILYDIALOG)

	zero-shot			fine-tuning			
Emotion	Precision	Recall	F1	Precision	Recall	F1	
neutral	=	_	_	=	_	_	
happiness	0.97	0.77	0.86	0.98	0.90	0.94	
surprise	0.53	0.16	0.25	0.48	0.85	0.61	
sadness	0.57	0.54	0.55	0.74	0.87	0.80	
anger	0.80	0.53	0.64	0.73	0.79	0.76	
disgust	0.39	0.19	0.26	0.71	0.11	0.19	
fear	0.26	0.82	0.40	0.72	0.76	0.74	
macro-average	0.59	0.50	0.49	0.73	0.71	0.67	
std	0.26	0.28	0.24	0.16	0.30	0.26	
micro-F1 (Ours)	-	-	0.75	-	-	0.86	
micro-F1 [11]	=	-	0.41	-	=	-	
micro-F1 [5]	=	-	0.43	=	=	-	

TABLE VI RECOGNITION PERFORMANCE WITH AN ALTERED LABEL (IEMOCAP)

Emotion	fine-tuning					
Emotion	Precision	Recall	F1			
neutral	0.68	0.57	0.62			
sad	0.86	0.68	0.76			
anger	0.45	0.72	0.55			
happiness	0.41	0.45	0.43			
frustrated	0.58	0.64	0.61			
excited	0.68	0.62	0.65			
macro-average	0.61	0.61	0.60			
std	0.17	0.09	0.11			
weighted-f1	-	-	0.62			

TABLE VII RECOGNITION PERFORMANCE AFTER FINE-TUNING ON A SINGLE DATASET (MELD)

Emotion	fine-tuning on a single dataset			fine-tuning on a mixed dataset		
Linotion	Precision	Recall	F1	Precision	Recall	F1
neutral	0.78	0.82	0.80	0.79	0.67	0.73
sadness	0.52	0.41	0.46	0.43	0.48	0.45
anger	0.61	0.64	0.62	0.53	0.59	0.56
disgust	0.52	0.22	0.31	0.37	0.28	0.32
fear	0.39	0.28	0.33	0.33	0.33	0.33
joy	0.62	0.65	0.64	0.61	0.62	0.62
surprise	0.60	0.57	0.59	0.47	0.69	0.56
macro-average	0.58	0.51	0.53	0.50	0.52	0.51
std	0.12	0.22	0.18	0.16	0.17	0.15
weighted-f1 (Ours)	-	-	0.68	-	-	0.63

#### IV. DISCUSSION

In this study, we evaluated ChatGPT's ability to recognize emotions associated with text. Within the context of the ERC task, ChatGPT demonstrated a commendable performance, surpassing chance levels and aligning with prior research findings. While fine-tuning generally improved recognition, we observed remaining variability in performance across different emotion labels and datasets.

The variability in recognition performance can be attributed to biases in the training data and the labeling methodologies used [12]. In prior research that attempted to generate datasets with emotion labels using ChatGPT, Koptyra et al. [3] reported variability in the number of generated labels. This observation aligns with our result and both suggest that the inherent bias in ChatGPT may influence emotion analysis. In the context of emotion recognition, biases towards certain label classes have been reported in pioneering language models [20]. Consequently, it is of utmost importance for researchers to understand and account for the intrinsic biases of language models when applying them to emotion analysis.

LLMs are praised for their advanced linguistic skills, making them an appealing solution for various applications, including mental health ([5]–[7], [21], [22]; see [23] for the survey). Given the potential biases reflected in emotion recognition, it would be prudent for researchers and practitioners to exercise caution when applying applying these models in sensitive domains like mental health. Over-reliance on these models without thorough evaluation might inadvertently propagate biases or yield misleading results.

Even within datasets such as MELD and EmoryNLP, which are based on the same television series, there was a notable

variance in performance, with particularly low performances for "powerful" and "peaceful." We hypothesize that one reason these emotion labels are not adequately associated with the text might be due to biases within the knowledge base of ChatGPT. Figure 4 presents examples of ChatGPT's responses when inquired about commonly used emotion labels. The responses from ChatGPT encompass widely accepted emotional definitions, such as Ekman's basic emotions [24] and Plutchik's wheel of emotions [25]. While the input-output mechanisms of LLM are black-box in nature, the fact that the model produces such knowledge suggests that for ChatGPT, associating these emotions with the input text might be relatively easier than associating less commonly used labels. These experimental results, alongside the sensitivity of the labels to the text, highlight the importance of label selection.

## A. Limitation of this Study

While our study sheds light on several aspects of emotion recognition using ChatGPT, there are certain limitations that should be acknowledged:

- 1) Inconsistencies in labeling criteria: Different datasets might have similar labels but with varied labeling criteria. This could account for the observed discrepancies in our results, a sentiment discussed in prior research [11].
- 2) **Potential for Further Optimization:** It is noteworthy that our experimental results might not fully reflect the optimal performance of ChatGPT for the corresponding tasks. As our results underscore the model's sensitivity to specific label texts, an adjustment of prompt could enhance its performance. We believe that through more refined prompt engineering and careful selection of context examples, the performance of ChatGPT can be further enhanced.
- 3) Model Specificity: Our study focused exclusively on the evaluation of ChatGPT, a representative large language model. Whether similar results would be observed when using other models remains undetermined.

In future research, a broader study that encompasses various linguistic and cultural contexts, as well as a wider range of emotion labels and models, could offer a more holistic perspective on LLMs' strengths and shortcomings in emotion recognition.

#### V. CONCLUSION

Our exploration of ChatGPT's capabilities in emotion recognition has illuminated both its potential and pitfalls. While the model demonstrates impressive proficiency, especially when fine-tuned, it is susceptible to inconsistencies across various emotion labels and datasets. These variations might be influenced by underlying biases, highlighting the importance of meticulous dataset and label selection. As the allure of deploying models like ChatGPT in critical domains grows, it is paramount to approach their integration with caution and diligence, ensuring their application remains reliable and unbiased in real-world contexts.

(User) In the context of emotion recognition research, what are the commonly used emotion labels?

(ChatGPT) In emotion recognition research, there are several commonly used emotion labels that researchers use to categorize and study human emotional states. These labels are typically based on various emotion theories and models. Some of the most common emotion labels include:

- 1) Basic Emotions: These are often considered the core, universally recognized emotions proposed by Paul Ekman. They include:
  - Happiness
  - Sadness
  - Anger

  - Disgust Surprise
- 2) Emotion Intensity: Researchers may also measure the intensity or strength of emotions, ranging from very weak to very strong.
- Valence-Arousal Model: This model categorizes emotions based on their valence (positive, negative, or neutral) and arousal (intensity or activation). Common labels include:
  - Positive emotions (e.g., joy, love)
  - Negative emotions (e.g., anger, sadness)
  - Neutral emotions (e.g., boredom, indifference)
- 4) Plutchik's Wheel of Emotions: Robert Plutchik's model suggests that there are eight primary emotions, which can be combined to form more complex emotional states. These primary emotions include:
  - Joy Sadness

  - Anger Fear
  - Trust
  - Disgust
  - Surprise
  - Anticipation

.. It's important to note that the choice of emotion labels can vary depending on the specific goals of the research and the theoretical framework being used. Additionally, advances in emotion recognition technology and machine learning have led to more fine-grained and automated approaches for detecting and categorizing emotions, which may use a combination of these labels or even entirely different methods.

Fig. 4. An example of the output of ChatGPT regarding commonly used emotion labels

## ACKNOWLEDGMENT

This study was conceptualized, conducted, and written by the authors, and an AI (OpenAI's GPT-4 model) was used for proofreading.

#### REFERENCES

- [1] K. Terada, A. Yamauchi, and A. Ito, "Artificial emotion expression for a robot by dynamic color change," in 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 314-321, IEEE, 2012.
- [2] D. Löffler, N. Schmidt, and R. Tscharn, "Multimodal expression of artificial emotion in social robots using color, motion and sound," in Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 334-343, 2018.
- [3] B. Koptyra, A. Ngo, Ł. Radliński, and J. Kocoń, "Clarin-emo: Training emotion recognition models using human annotation and chatgpt,' in International Conference on Computational Science, pp. 365-379, Springer, 2023.
- [4] Y. Zhang, Z. Wang, and J. Shang, "Clusterllm: Large language models as a guide for text clustering," arXiv preprint arXiv:2305.14871, 2023.
- [5] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis," arXiv preprint arXiv:2304.03347, 2023.
- [6] B. Lamichhane, "Evaluation of chatgpt for nlp-based mental health applications," arXiv preprint arXiv:2303.15727, 2023.
- [7] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt," arXiv preprint arXiv:2303.03186, 2023.
- C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is chatgpt a general-purpose natural language processing task solver?," arXiv preprint arXiv:2302.06476, 2023.

- [9] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert," arXiv preprint arXiv:2302.10198, 2023.
- [10] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, et al., "Chatgpt: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.
- [11] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is chat-gpt equipped with emotional dialogue capabilities?," arXiv preprint arXiv:2304.09582, 2023.
- [12] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, "Decoding chatgpt: a taxonomy of existing research, current challenges, and possible future directions," *Journal of King Saud University-Computer and Information Sciences*, p. 101675, 2023.
- [13] OpenAI, "Chatgpt." https://openai.com/blog/chatgpt. Accessed: 2023-08-05.
- [14] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
- [15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
- [16] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," arXiv preprint arXiv:1810.02508, 2018.
- [17] S. M. Zahiri and J. D. Choi, "Emotion detection on tv show transcripts with sequence-based convolutional neural networks," in Workshops at the thirty-second agai conference on artificial intelligence, 2018.
- [18] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," arXiv preprint arXiv:1710.03957, 2017.
- [19] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [20] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, 2022
- [21] Z. Elyoseph, D. Hadar-Shoval, K. Asraf, and M. Lvovsky, "Chatgpt outperforms humans in emotional awareness evaluations," Frontiers in Psychology, vol. 14, p. 1199058, 2023.
- [22] K. Uludag, "Testing creativity of chatgpt in psychology: Interview with chatgpt," Available at SSRN 4390872, 2023.
- [23] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria, "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics," arXiv preprint arXiv:2310.05694, 2023.
- [24] P. Ekman, "Are there basic emotions?," 1992.
- [25] R. Plutchik, "A general psychoevolutionary theory of emotion," in Theories of emotion, pp. 3–33, Elsevier, 1980.