Emotional Cause Analysis Based on In-Context Learning of Large Language Models

Kongqiang Wang[1*], Qingli Tan[2], and Peng Zhang[1]

[1]School of Information Science and Engineering, Yunnan University, Kunming, Yunnan, China

[2]College of Ecology and Environment, Yunnan University, Kunming, Yunnan, China

*Corresponding author: wangkongqiang60@gmail.com

Abstract

The ability to understand emotions is an essential component of human-like artificial intelligence, as emotions greatly influence human cognition, decision making, and social interactions. In addition to emotion recognition in conversations, the task of identifying the potential causes behind an individual's emotional state in conversations is of great importance in many application scenarios. The main content of our research is Multimodal Emotion Cause Analysis in Conversations (MECAC), which aims at extracting all pairs of emotions and their corresponding causes from conversations. Under different modality settings, it consists of two specific circumstances: Textual Emotion-Cause Pair Extraction in Conversations (TECPE) and Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE). For this shared study, the main dataset used is a multimodal emotion cause dataset ECF 2.0 sourced from the sitcom *Friends*. This dataset contains 1,715 conversations and 16,720 utterances, where 12,256 emotion-cause pairs are annotated at the utterance level, covering three modalities (language, audio, and vision). We conducted follow-up research on this benchmark dataset, using mainstream large language models (LLMs) such as GPT3.5, GPT4V, Llama2, Llama3 and Flan-T5. We achieved remarkable results for this challenging task in the multimodel emotion cause analysis field. For Textual Emotion-Cause Pair Extraction in Conversations (TECPE), we achieved an emotional cause extraction result of ACC_cause 0.5708 and F1_cause 0.3243. For Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE), we achieved an emotional cause extraction result of w-avg. Strict F1 0.3982 and Strict F1 0.4017.

Index Terms— Emotional Cause Analysis, Knowledge Reasoning, Large Language Model (LLM), Flan-T5, Llama, GPT

1 Introduction

Understanding emotions is crucial to achieve human-like artificial intelligence, as emotions are intrinsic to humans and significantly influence our cognition, decision-making, and social interactions. Conversation is an important form of human communication and contains a large number of emotions. Furthermore, given that conversation in its natural form is multimodal, many studies have explored multimodal emotion recognition in conversations (ERC), using language, audio and vision modalities [1].

However, emotion recognition alone is not sufficient to fully understand the intricacies of human emotions. Emotion cause analysis (ECA), the process of identifying the potential causes behind an individual's emotion state, has broad application scenarios such as human-computer interaction, commerce customer service, empathetic conversational agents, and automatic psychotherapy. For example, conversational agents equipped with emotion cause analysis can better understand the user's emotional state, offer empathetic responses, and provide more personalized services. By identifying the cause of the emotional state of a patient, a psychotherapy system can provide more accurate and customized treatments. Emotion cause analysis (ECA) has gained increasing attention both in academic and practical fields [2]. To promote research in this direction, we join in Multimodal Emotion Cause Analysis in Conversations (MECAC) shared study task. This task consists of two specific circumstances: Textual Emotion-Cause Pair Extraction in Conversations (TECPE), focuses on extracting emotion and textual cause spans solely based on text content. Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE), involves extracting emotion-cause pairs at the utterance level considering three modalities: language, audio, and video.

For this shared study task, Wang et al. [13] provide a multimodal emotion cause dataset ECF 2.0 sourced from the sitcom *Friends*. This dataset contains 1,715 conversations and 16,720 utterances, where 12,256 emotion-cause pairs are annotated at the utterance level, covering three modalities (language, audio, and vision). Specifically, in their preliminary work [3]. They have constructed a benchmark dataset, Emotion-Cause-in-Friends (ECF 1.0), which contains 1,374 conversations and 13,619 utterances. On this basis, they have furthermore annotated an extended test set as the evaluation data and provided the span-level annotations of emotion causes within the textual modality.

We proposing numerous well-designed pipeline systems. Moreover, we applied advanced large language models (LLMs) for emotion cause analysis and achieved promising results. These large language models (LLMs) include GPT3.5,

GPT4V, Llama2, Llama3 and Flan-T5.

2 Related Works

2.1 Emotion Recognition in Conversations

Emotion Recognition in Conversations (ERC) has been extensively studied as a fundamental task for understanding human emotions in conversation scenarios. Early studies mainly focused on textual information and modeled contextual dependencies among utterances using recurrent neural networks and attention mechanisms. To better capture conversational dynamics, researchers further introduced graph-based approaches to model speaker interactions and contextual relationships across utterances. These methods significantly improved emotion classification performance by leveraging conversational structure and contextual cues.

With the development of multimodal learning, recent studies have incorporated acoustic and visual signals to enhance emotion recognition performance. Multimodal ERC approaches utilize complementary information from language, speech, and facial expressions, demonstrating that multimodal fusion provides richer emotional representations than unimodal methods. Despite these advances, ERC primarily focuses on emotion classification and does not explicitly identify the underlying causes of emotions.

2.2 Emotion Cause Analysis

Emotion Cause Analysis (ECA) aims to identify the underlying reasons that trigger emotional states, which provides deeper interpretability compared to emotion recognition alone. Early research mainly focused on emotion cause extraction in single documents, where rule-based methods and traditional machine learning models were first explored. Later approaches adopted neural architectures such as convolutional neural networks and attention-based models to capture semantic relationships between emotions and their causes.

More recent studies extended emotion cause extraction to conversational settings, introducing Emotion-Cause Pair Extraction in Conversations (ECPE), which aims to jointly identify emotions and their corresponding causes. These methods typically model interactions between emotion and cause clauses using multi-task learning or joint prediction frameworks. However, most existing approaches rely solely on textual information and do not fully exploit multimodal signals available in real-world conversations.

2.3 Multimodal Emotion Cause Analysis

Given that human communication naturally involves multiple modalities, recent work has explored multimodal emotion cause analysis by incorporating textual, acoustic, and visual information. Multimodal approaches aim to leverage complementary cues from speech prosody, facial expressions, and linguistic content to improve emotion and cause prediction.

To facilitate research in this area, benchmark datasets such as Emotion-Cause-in-Friends (ECF) have been proposed, providing multimodal conversational data with annotated emotion-cause pairs. These datasets enable the study of Multimodal Emotion Cause Pair Extraction (MECPE), which requires modeling complex cross-modal interactions and temporal dependencies. Despite recent progress, multimodal emotion cause analysis remains challenging due to modality heterogeneity, complex emotion-cause relationships, and limited annotated data.

2.4 Large Language Models for Emotion Understanding

Large language models (LLMs) have recently demonstrated strong capabilities in natural language understanding and reasoning tasks. Their ability to capture contextual semantics and perform knowledge-driven inference makes them promising for emotion analysis tasks. Recent studies have explored the application of LLMs to emotion recognition, sentiment analysis, and affective reasoning through prompt-based learning and instruction tuning.

Furthermore, multimodal large models extend these capabilities by integrating visual and auditory inputs, enabling cross-modal reasoning and improved contextual understanding. Despite their success in many NLP tasks, the use of LLMs for multimodal emotion cause analysis in conversations remains underexplored, particularly for jointly extracting emotion-cause pairs.

In this work, we investigate Multimodal Emotion Cause Analysis in Conversations (MECAC) under both textual and multimodal settings. We design effective pipeline frameworks and explore mainstream large language models for emotion-cause reasoning. Our approach evaluates multiple LLMs on the ECF 2.0 benchmark and demonstrates promising performance for both Textual Emotion-Cause Pair Extraction (TECPE) and Multimodal Emotion-Cause Pair Extraction (MECPE), providing new insights into leveraging large models for multimodal affective computing tasks.

3 Dataset

3.1 Data Analysis

Sitcoms come with real-world-inspired inter-human interactions and usually contain more emotions than other TV series or movies. Based on the famous American sitcom *Friends*, Poria et al. [4] constructed the multimodal conversational dataset MELD by extracting audiovisual clips corresponding to the scripts of the source episodes and annotating each utterance with one of six basic Ekman emotions (*Anger*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise*) or *Neutral*. MELD has recently become a widely used benchmark for multimodal emotion recognition in conversations (ERC). In preliminary work, Wang el al. [3] chose MELD as the data source and further annotated the causes given emotion annotations, thereby constructing

the ECF 1.0 dataset. For this shared study, they release the entire ECF 1.0 dataset as a training set and additionally create a test set as evaluation data, which is also sourced from *Friends*. They employ three graduate students involved in the annotation of the ECF 1.0 dataset to annotate the extended test set. Given a multimodal conversation, they first need to annotate the speaker and emotion category for each utterance, and then further annotate the utterances containing corresponding causes for each non-neutral emotion. If the causes are explicitly expressed in the text, they should also mark the textual cause spans. After annotation, they determine the emotion categories and cause utterances by majority voting, and take the largest boundary (i.e., the union of the spans) as the gold annotation of the textual cause span. If disagreements arise, another expert is invited for the final decision.

The following tables presents a comparison of existing emotion cause analysis (ECA) datasets and ECF datasets, refer Table 1.

Table 1: Comparison of existing emotion cause analysis (ECA) datasets.

Dataset	Modality	Scene	Content
Emotion-Stimulus[5]	T	–	2,414 s[1]
ECE Corpus[6]	T	News	2,105 d[1]
NTCIR-13-ECA[7]	T	Fiction	2,403 d[1]
Weibo-Emotion[8]	T	Blog	7,000 p[1]
REMAN[9]	T	Fiction	1,720 d[1]
GoodNewsEveryone[10]	T	News	5,000 s[1]
RECCON-IE[11]	T	Conv	665 u[1]
RECCON-DD[11]	T	Conv	11,104 u[1]
ConvECPE[12]	T,A,V	Conv	7,433 u[2]
ECF 1.0[13]	T,A,V	Conv	13,619 u[2]
ECF 2.0	**T,A,V**	**Conv**	**16,720 u**[2]

Note: T, A, and V refer to text, audio, and video. Blog and Conv represent microblog and conversation, and s, d, p and u denote sentence, document, post and utterance.
[1] text modality dataset.
[2] multimodal dataset (text, audio and video).

Table 2: Statistics of ECF 1.0, Extended Test, and ECF 2.0 datasets.

Items	**ECF 1.0**	**Extended Test**	**ECF 2.0**
Conversations	1,374	341	1,715
Utterances	13,619	3,101	16,720
Emotion (utterances)	7,690	1,821	9,511
TECPE			
Emotion (utterances) with causes	6,761	1,626	8,387
Emotion-cause (span) pairs	9,284	2,256	11,540
MECPE			
Emotion (utterances) with causes	7,081	1,746	8,827
Emotion-cause (utterance) pairs	9,794	2,462	12,256

3.2 Dataset Statistic

In preliminary work, they have already constructed the ECF 1.0 dataset that contains 1,374 conversations and 13,619 utterances. Furthermore, they have annotated an extended test set specifically for this study work evaluation, which together with ECF 1.0 constitutes the ECF 2.0 dataset that contains 1,715 conversations and 16,720 utterances.

Table 3: Statistics of annotating each utterance with one of six basic emotions in ECF 2.0 training datasets.

	cause situation[1]			self-causal probability[2]		
Project(number)	self-causal	non-self-causal	no-cause	self-causal	isolated	not-isolated
anger(1615)	886	544	185	0.5486	0.5090	0.4910
joy(2301)	1716	443	142	0.7458	0.7080	0.2920
sadness(1147)	756	309	82	0.6591	0.6746	0.3254
surprise(1840)	1049	699	92	0.5701	0.6168	0.3832
disgust(414)	298	87	29	0.7198	0.6275	0.3725
fear(373)	187	107	79	0.5013	0.7701	0.2299

Note: By analyzing each utterance annotated with one of six basic emotions in the ECF 2.0 training dataset, most cause cases have self-causes. [1] Number of six basic emotion utterances, divided into self-causal, non-self-causal and no-cause. [2] Probability that six basic emotion utterances are self-causal and whether the calculation is isolated.

In Table 1, we compare emotion cause analysis (ECA) existing dataset with the related datasets for experiment, in terms of modality, scene, and size. It is evident that ECF 2.0 is currently the largest available emotion cause dataset.

Table 2 and Table 3 presents the detailed statistics of ECF 2.0 dataset for the two specific circumstances. It can be seen that, in the entire ECF 2.0 dataset, 56.88% of the utterances are labeled with one of the six basic Ekman emotions, 92.81% of the emotion utterances have corresponding cause utterances, and 88.18% of the emotion utterances are annotated with textual cause spans.

In addition, as shown in Figure 1 and Figure 2, the newly annotated test set is basically consistent with the original ECF 1.0 dataset in terms of conversation length and emotion categories distribution.

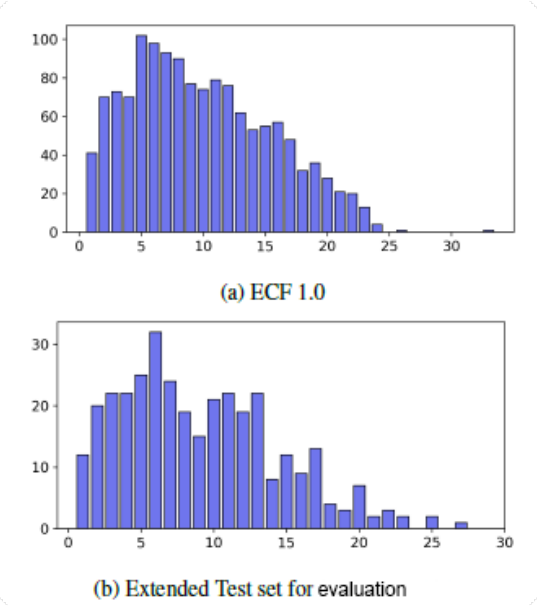

Figure 1: The distribution of conversation lengths. The horizontal axis represents the number of utterances, and the vertical axis represents the number of conversations.

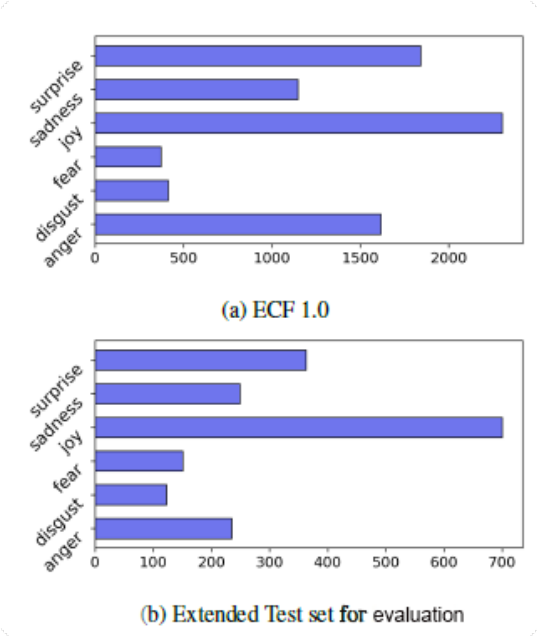

Figure 2: The distribution of emotion categories. The horizontal axis represents the number of utterances, and the vertical axis represents emotion categories.

3.3 Evaluation Metrics

We evaluate the emotion-cause pairs of each emotion category with F1 scores separately and further calculate a weighted average of F1 scores across the six emotion categories, denoted as "w-avg. F1". Specifically, for Textual Emotion-Cause Pair Extraction in Conversation (TECPE), which involves the textual cause span, we adopt two strategies to determine whether the span is extracted correctly:

• **Strict Match**: A predicted span is regarded as correct if it's the same as one of the annotated spans;

• **Proportional Match**: Calculate the overlap proportion of the predicted span and the annotated one.

The evaluation metrics for the two strategies are "w-avg. S. F1" and "w-avg. P. F1", respectively. Taking into account the complexity of Textual Emotion-Cause Pair Extraction in Conversation (TECPE), we choose "w-avg. S. F1" as the main metric for the ranking. Then, for Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE), we consider the complexity of multimodality (text, audio ,video). We no longer consider the issue of text span, but only need to consider the emotional cause matching.

3.4 Metrics Equations

3.4.1 Textual Emotion-Cause Pair Extraction in Conversations (TECPE)

Here we explain how the evaluation metrics are calculated.

Given a conversation, each emotion-cause pair p_i in this subtask should contain five elements: index of emotion utterance eu_i, emotion category ec_i, index of cause utterance cu_i, start index of cause span ss_i, end index of cause span se_i, i.e, $p_i = [eu_i, ec_i, cu_i, ss_i, se_i]$. For the textual cause span, we adopt two strategies to determine whether the span is extracted correctly:

- **Strict Match**: the predicted span should be exactly the same as the annotated span.
- **Proportional Match**: considering the overlap proportion of the predicted span and the annotated one.

The evaluation script will output the results of multiple metrics, among which the main evaluation metric used for ranking is *w-avg. Strict F1*. The following explains the calculation method of *Strict F1*. The Precision, Recall, and F1 scores are calculated as follows:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{1}$$

where,

$$P = \frac{\sum CorrectPairs}{\sum PredictedPairs'} \tag{2}$$

$$R = \frac{\sum CorrectPairs}{\sum AnnotatedPairs'} \tag{3}$$

where $\sum$ means summing over all conversations in the dataset, *PredictedPairs* denotes the number of emotion-cause pairs predicted by the model, *AnnotatedPairs* denotes the number of annotated emotion-cause pairs, and *CorrectPairs* means the number of correct pairs in a conversation. A predicted pair is regarded as correct if its five elements are all the same as one of the annotated pairs. The content following explains the calculation method of *w-avg. Strict F1* (main). We first evaluate the emotion-cause pairs of each emotion category with *Strict F1* scores separately:

$$F_1^j = \frac{2 \times P^j \times R^j}{P^j + R^j} \tag{4}$$

where,

$$P^j = \frac{\sum CorrectPairs^j}{\sum PredictedPairs^j} \tag{5}$$

$$R^j = \frac{\sum CorrectPairs^j}{\sum AnnotatedPairs^j} \tag{6}$$

where *Pairs*^{j} denotes the number of pairs with emotion category *j*, this category $j \in \{$, disgust, fear, joy, sadness, surprise$\}$. Then we further calculate a weighted average of *Strict F1* scores across the six emotion categories:

$$F_1 = \sum_{j=1}^{6} w^j F_1^j \tag{7}$$

where w^j is the proportion of annotated pairs with emotion category *j*, i.e.,

$$w^j = \frac{\sum AnnotatedPairs^j}{\sum AnnotatedPairs}. \tag{8}$$

The following explains the calculation method of *Proportional F1*. Under the condition that the three elements $[eu_i, ec_i, cu_i]$ are the same, we match each predicted pair with one of the annotated pairs that has the maximum overlap proportion in terms of the cause span (if the predicted span overlaps with multiple annotated spans):

$$\text{overlap}_i = \begin{cases} \text{len}(ps_i \cap as_k), & \text{if } [eu_i, ec_i, cu_i] \text{ are correct} \\ & \text{and } ps_i \cap as_k \neq \emptyset, \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$k = \arg\max_t \frac{\text{len}(ps_i \cap as_t)}{\text{len}(as_t)} \tag{10}$$

where len(*) denotes the number of textual tokens, ps_i and as_k represent the cause span in the predicted pair pp_i and the annotated pair ap_k respectively. Then the *proportional* F1 is calculated based on the overlap length between the predicted span and the annotated span:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

where,

$$P = \frac{\sum \sum_i \text{overlap}_i}{\sum \sum_i \text{len}(ps_i)} \tag{12}$$

$$R = \frac{\sum \sum_i \text{overlap}_i}{\sum \sum_t \text{len}(as_t)} \tag{13}$$

where *i* and *t* denote the index of predicted pairs and annotated pairs in a conversation respectively.

The following explains the calculation method of *w-avg. Proportional F1*. Similar to the *w-avg. Strict F1*, the *w-avg. Proportional F1* is a weighted average of *Proportional F1* scores across the six emotion categories.

3.4.2 Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE)

Given a conversation, each emotion-cause pair in this task should contain three elements: index of emotion utterance eu_i, emotion category ec_i, index of cause utterance cu_i, i.e, p_i = $[eu_i, ec_i, cu_i]$. The main evaluation metric is the *w-avg. F1*. We also calculate the *micro F1* score. A predicted pair is regarded as correct if its three elements are all the same as one of the annotated pairs.

4 Methodology

In the field of natural language processing (NLP), emotion cause analysis (ECA) is a challenging task. It not only requires the model to identify the emotional states expressed in conversations or texts, but also accurately identify the specific causes of such emotions. This emotion cause ananlysis in conversation (ECAC) study work focuses on modeling the correlation between emotions and their causes in multimodal dataset. The data sources mainly come from real conversation clips of the American TV series *Friends*, including text, visual frames and voice cues. Traditional emotion recognition tasks mostly focus on emotion classification (such as anger, happiness, sadness, etc.), but this task goes a step further, requiring the model not only to identify emotion categories but also to find the causes of their generation in the context. This places extremely high demands on the model's language understanding ability, causal reasoning ability and multimodal fusion ability. Based on this challenging problem, we have proposed three different approaches to address this difficulty.

4.1 Three-Hop Reasoning

Concept of Three-Hop Reasoning: aspect extraction, opinion extraction, and reasoning on sentiment. Finally produced sentiment class [14]. The core issue of the research is: How can an Large Language Model (LLM) reason out the emotions expressed by a character in a specific context? How can we further identify the emotional cause that triggers this emotion from multiple rounds of conversations? How can emotional states and their causes be uniformly modeled to achieve joint prediction in one model? To solve these problems, we have proposed a methodology called Three-Hop Reasoning (THoR), emphasizing the gradual extraction of emotional cues in conversations and conducting multiple rounds of causal reasoning.

This method simulates the way humans understand the causality of emotions use Flan-T5 model[1] and divides the overall reasoning process into three steps: **Emotion State Extraction**. The input is the context containing several historical speeches before the target speech. The model first determines the type of emotion currently expressed by the target speaker. **Emotion Cause Extraction**. After knowing the emotional state, look for the source event or statement that triggered the emotion in the context. This usually involves identifying the specific content of other speakers or earlier statements to form causal connections. **Reasoning Revision**. Based on the first two steps, the model further reviews the reasoning path and combines causal information to correct the understanding of emotional states, thereby improving accuracy. This step can be regarded as a self-supervised knowledge correction process. The design of this three-hop structure reflects the process of starting from the state, reviewing the context, and correcting the judgment, enhancing the model's causal modeling ability in complex contexts. For an overall detailed summary, please refer to Figure 3.

The following is a introduction of model design and training methods. We used Google's open-source model Flan-T5[2] as the basic model and conducted two-stage training [18]:

• Phase one: Fine-tuning of the emotional state recognition task using the state subset of the ECAC dataset, the model focuses on learning the emotions directly expressed in the text.

• Phase two: Identification and correction of emotional causes. The model was fine-tuned for the second time using

[1]Flan-T5: https://arxiv.org/pdf/2210.11416

[2]Hugging Face: https://huggingface.co/google/flan-t5-base

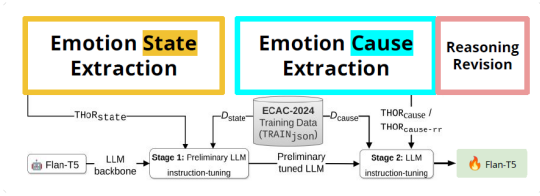

Figure 3: The situation of two-stage approach for tuning LLM (Flan-T5).

cause data, and the reasoning revision mechanism was added simultaneously to enhance its robustness. The training objective is to simultaneously predict the emotion category and the specific sentence or behavior that causes the emotion.

The two stages are trained on different task data respectively, but share the underlying encoder parameters to form a joint optimization framework. The preliminary fine-tuning experiment on Dev dataset of emotion state, see Figure 4.

The following is a introduction of experimental design and result analysis. The evaluation was conducted on the development dataset provided by ECF 2.0[3], with the main focus indicator being the F1 score. The key experimental findings are as follows:

• Baseline method (PROMPT): The prompt-based solution has performed well in cause dataset.

• Three-hop reasoning method (THoR): The THoR-cause method performs slightly better than the prompt baseline.

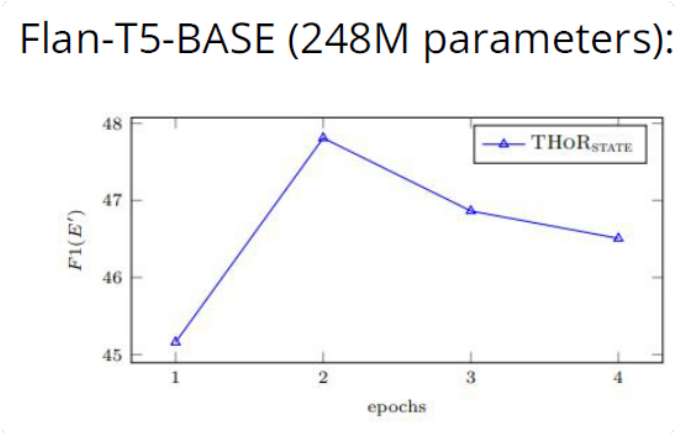

Figure 4: Result analysis of the preliminary fine-tuning of *Flan-T5$_{base}$* on D_{state} *dev* using *THoR$_{STATE}$* technique per epoch by $F_1(E')$.

• Reasoning revision strategy (RR): After the introduction of reasoning revision, the F1 score on the development set increased by approximately 2.1%, showing a clear gain.

These results indicate that by refining the reasoning process and incorporating a feedback mechanism, the model [20] can grasp the complex relationship between emotions and their causal structure more accurately. For specific fine-tuning details, please refer to Figure 5.

[3]Hugging Face: https://huggingface.co/datasets/NUSTM/ECF

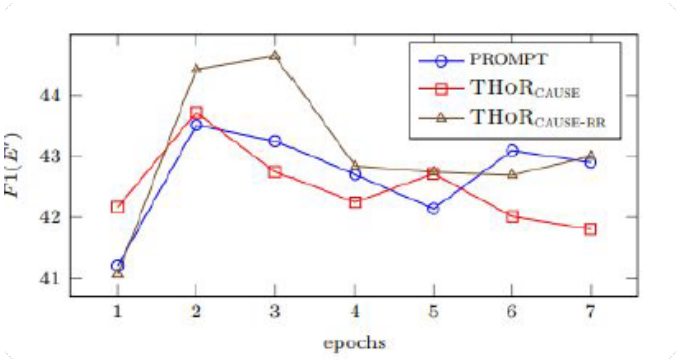

Figure 5: Flan-T5$_{\text{base}}$ fine-tuning results comparison by $F1(E')$ on $D_{\text{cause dev}}$ part per each epoch across fine-tuning techniques: PROMPT, THOR$_{\text{CAUSE}}$, and THOR$_{\text{CAUSE-RR}}$.

Spans correction algorithm implementation of the vocabulary-based spans-correction technique mentioned in Algorithm 1.

Algorithm 1 Emotion-cause prefixes correction for u_{text}

updated ← True
V_p' ← sorted V_p by decreased entry lengths in words
while while $V_p' \neq \varnothing$ or updated **do**
    updated ← False
    $u_{\text{text}} \leftarrow u_{\text{text}}$ ▷ Modified version of u_{text}
    for $v_p \in V_p'$ **do**
        if if u_{text} ends with v_p then **then**
            u_{text} ← part of u_{text} before v_p
            updated ← True
            break
        end if
    end for
end while

4.2 Fine-tuned Llama-2 and Llama-3

We perform instruction fine-tuning of the Llama 2 large language model (LLM), an open-source model developed by GenAI, Meta [15]. From the three variants with 7, 13, and 70 billion parameters, we use the 7 billion parameter model due to resource constraints, albeit the performance of this model achieves state-of-the-art results on various downstream NLP tasks compared to other models of similar sizes. In addition, we use the Llama 2-chat version of the model, which is optimized for dialogue use cases as it aligns with our task. In our approach, we use Llama2 API for prompt engineering. Through zero-shot prompting, we select optimal prompts for emotion identification and cause prediction. We observed that treating these two tasks separately resulted in better model output. This approach involves first identifying the emotions of all utterances in the conversation. We then add these emotion labels to the conversation and prompt the model to predict the causes for each emotion utterance. Consequently, we perform supervised fine-tuning of two separate Llama 2 models

for these tasks. Although this increases the inference time, the significant performance gains outweigh the introduced latency. We treat both tasks as conditional generation, where the model generated the emotion label in the first case and the cause list in the second case on the condition of given the prompt. Detailed explanations of these approaches are provided in the following sections.

The following is a deep analysis of the Llama 3.2 large language model (LLM). The model categories include small and medium-sized visual models (11B and 90B), as well as lightweight text models (1B and 3B), which are suitable for edge devices and mobile devices. They are all highly innovative. The innovation of the visual model lies in its support for image reasoning for the first time. The 11B and 90B models combine the image encoder with the language model through adapters, achieving the alignment of text and images. In the later training optimization, methods such as supervised Fine-tuning (SFT) and Preference Optimization (DPO) are adopted to enhance the model's understanding and reasoning abilities on image and text prompts. The following is a introduction to the Llama 3 large language model (LLM). Meta has developed and released the Meta Llama 3 model series, which includes pre-trained and instruct-tuned generated text models with parameters of 8B and 70B respectively. The release time is April 18, 2024. The input of the model is limited to text input only, and the output of the model is only generated text and code.

For the Llama 3 environment configuration, model download and model inference, the versions available for us on Hugging Face are meta-llama/Meta-Llama-3-8B[4] and meta-llama/Meta-Llama-3-8B-Instruct[5]. Due to the limitations of the experimental environment, with only one NVIDIA GeForce RTX 3090 (24G) and one Nvidia A100 40GB GPU graphics units available, we selected the meta-llama/Meta-Llama-3-8B-Instruct, meta-llama/Meta-Llama-2-13B[6], and meta-llama/Meta-Llama-2-7B[7] model version as the object for fine-tuning the large language model (LLM). The main inference frameworks of Llama are inference based on transformers and model inference based on VLLM. Since VLLM only supports the Linux operating system and has incompatibility on the Windows operating system, we ultimately choose the transformer inference framework as the experiment inference framework. These are all the best decisions made based on objective reality conditions. In fact, it have also proved that the experimental effect of our choice of Llama 3 (8B) is much better than that of Llama 2 (7B).

To perform emotion recognition, we create a dataset where each sample includes an utterance u_j from one of the N conversations D for which the large language model (LLM) needs to output the emotion label. We incorporate the entire conversation D_i along with speaker information as context in our prompt. This contextual information enhances the model's understanding of the flow of emotions within the conversation. The instruction I_j^e which gave the best results is given along with detailed prompt examples. The prompt consists of the instruction I_j^e and the context C_j for utterance u_j.

$$Prompt_j = (C_j, I_j^e) \tag{14}$$

Using this prompt as the input and the corresponding true emotion label y_j^e, we perform supervised fine-tuning of a Llama 2-7b, Llama 2-13b and Llama 3-8b model. This enables the model to ultimately achieve the prediction results as shown in the following formula. Greatly improve the model's prediction of the speaker's emotional state. Among them, θ represent the model parameters.

$$\hat{y}_j^e = \mathbf{Llama}_e(Prompt_j, \theta) \tag{15}$$

We use a quantized version of the model due to memory limitations and perform Quantized Low-Rank Adaptation (QLoRA) [19] as a parameter-efficient fine-tuning technique. The overall execution process depends on Algorithm 2.

4.3 In-Context-Learning GPT

Our final approach tackles Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE) by obtaining conversation-level video captions using the GPT-4Vision model by OpenAI [16]. For emotion prediction, we retrieve a semantically similar conversation from the training set whose emotion annotations are explained as demonstration examples in the prompt for the GPT-3.5 model. For each predicted emotional utterance, we perform cause prediction within a context window around the emotional utterance. Due to the complex nature of the task, we leverage in-context-learning [17] by retrieving similar context windows from the training set whose cause annotations are explained as demonstration examples in the prompt for the GPT-3.5 model. We discuss each step in the subsequent sections.

GPT-4V has the capability to process video sequences [21]. In our approach, we extract conversation-level captions from the videos. However, due to rate limits and the costs considerations, we use a compact image representation for each video associated with the utterances of a conversation. Therefore, these image sequences serve as input to the GPT-4V model, generating a description for the entire conversation. For an utterance, we sample nine equidistant frames across its video length. These frames aim to capture the dynamics of the whole video. We arrange these frames in a 3 × 3 grid, following a row-major order. Additionally, we include the speaker text below the grid to provide further context to GPT-4V. GPT-3.5 tends to be uncontrollable when performing zero-shot recognition of emotions in conversations outputting emotions that are not a valid category of labels [22]. To guide and control the process, we leverage in-context learning (ICL) by retrieving a conversation from the training set whose emotions are already annotated. The emotions in these conversations are explained by GPT- 3.5. This retrieved conversation and its explanation serve as a demonstration for GPT-3.5 to learn from, enabling it

[4]Hugging Face: https://huggingface.co/meta-llama/Meta-Llama-3-8B

[5]Hugging Face: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[6]Hugging Face: https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[7]Hugging Face: https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

to recognize emotions in conversations more accurately. In addition, the prompt template includes the video caption as part of the input.

Following the prediction of emotions, we predict the causes for each emotion-labeled utterance within a context window around that utterance. The bounds of the context window are given, see Table 4. The bounds were informed by the distribution of the majority of relative positions of causes in the training, see Figure 6.

Table 4: The bounds of the context window at each directional position.

Position	Previous	Next
Beginning	0	2
End	5	0
Middle	5	2

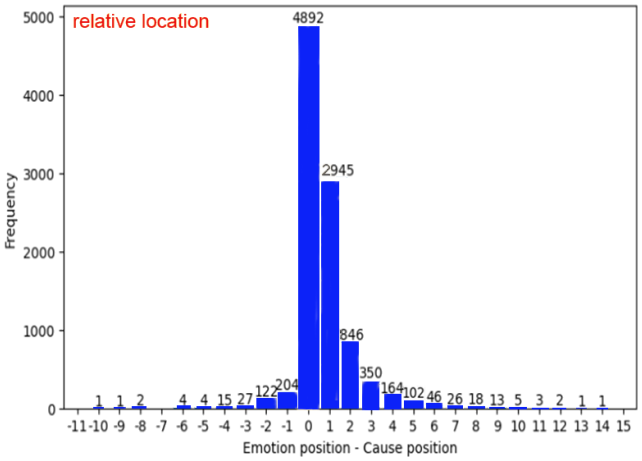

Figure 6: The relative position of emotion utterance and cause utterance. Among them, the vertical axis represents the number of emotion cause pairs, and the horizontal axis represents the difference between the emotion utterance position and the cause utterance position.

For predicting the causes of an utterance with emotion e within a given context window c, we retrieve context windows containing utterances with the same emotion e and position p that exhibit semantic similarity to c. This retrieval is accomplished through the Euclidean distance comparison of text-embedding-ada-002 embeddings derived from the training data. The retrieved conversation's causes are explained by GPT-3.5. Learning from the explained retrieved-context windows, cause prediction on c can be performed by GPT-3.5. Video captions are also included in the prompt. In our final approach, we perform a post-processing step where we add the emotion-labeled utterance as its own cause which we call self-causes after getting the causes. This operation gives significant performance boosts as a majority of the causes are self-causes as pointed out in Table 3. For a specific description of the GPT-based method, please refer to Algorithm 3. A general description of the GPT-based method, see Table 5.

Table 5: A general description of the steps of the GPT-based method used in ECF 2.0 training dataset and test dataset.

`+Video Captioning+`	Capture the dynamics of the whole video to become frames in a 3 × 3 grid following a row-major order, include the speaker text below the grid to provide further context to GPT-4V to generate video caption.
`+Emotion Recognition+`	The conversations retrieved from the training set and their explanations are used as demonstrations for GPT 3.5 learning, enabling it to recognize emotions in the conversations. Its prompt templates include video captions.
`+Cause Prediction+`	Predict the causes of each emotion utterance in the context window surrounding the utterance. Retrieve a window in which the context contains utterance with semantic similarity of the same emotion e and position p obtained from the training data. From the explained retrieval context window, GPT-3.5 can be used to predict the causes of conversations with labeled emotions. Video captions are also included in the prompt.
`+Post-Processing+`	We perform a post-processing step where we add the emotion-labled utterance as its own cause which we call self-causes.

5 Experiments

5.1 Results on Three-Hop Reasoning

We present a Chain-of-Thought (CoT) methodology aimed at fine-tuning LLM for emotion state prediction and cause extraction. We consider the problem of emotion cause analysis in conversations as a context-based problem with the mentioned utterance that causes emotion towards the last utterance in context. We devise the Chain-of-Thought (CoT) approach for emotion causes extraction and propose a reasoning revision methodology aimed at imputing the speaker emotion to support the decision on which utterance caused emotion. The Chain-of-Thought (CoT) represent a Three-Hop Reasoning approach prior known as THoR. We apply this approach to fine-tune LLM and predict emotion state of the mentioned utterance and the last utterance. The THoR_state model is less effective in emotion recognition compared with the effect of the prompt_state method on the validation set, see Table 6. Emotion caused by mentioned utterance towards the last utterance in context. The THoR_cause model is more effective in extracting emotional causes compared with the effect of the prompt_cause method on the validation set, see Table 7. We experiment with the Flan-T5-base (250M) model fine-tuning using resources provided by train dataset. The application of CoT with reasoning revision allows us to improve the results by 2.8% F1-measure compared to prompt-based tuning. This method is effective in emotion recognition and the extraction of emotional causes. Specific effect analysis see table 8.

5.2 Results on Fine-tuned Llama-2 and Llama-3

We conduct error analysis for the output of emotion recognition using these approaches. The performance of Llama-2-7B is slightly poor where the model predicts the label neutral for almost all utterances. On adding the conversational context, the model can identify the emotional nuances better, yet often

Table 6: Comparison of different engine the source of the prompts in on speakers emotion states extraction.

Approach	Epochs	Acc	F1	Mode
THoR_state	**0**	**59.729**	**43.757**	**valid**
-	1	55.729	42.476	valid
-	2	55.729	43.257	valid
-	3	57.220	42.693	valid
prompt_state	0	57.763	43.323	valid
-	1	58.712	44.949	valid
-	2	57.966	45.603	valid
-	**3**	**59.729**	**45.687**	**valid**

Table 7: Comparison of different engine the source of the prompts in on speakers emotion cause extraction.

Approach	Epochs	Acc	F1	Mode
THoR_cause	0	76.689	42.294	valid
-	**1**	**78.018**	**42.794**	**valid**
-	2	77.436	41.454	valid
-	3	77.741	42.141	valid
prompt_cause	**0**	**76.163**	**42.049**	**valid**
-	1	78.212	41.523	valid
-	2	77.215	41.326	valid
-	3	77.962	41.911	valid

predicts joy or surprise for neutral. Instruction fine-tuning significantly boosts performance where the model can now differentiate distinct emotions. The performance on disgust and fear is low due to the class-imbalance problem. In our test subset, the support of disgust and fear is only a little. The accuracy rate of these methods in emotion recognition as shown in Table 9. The specific loss effects and post-processing methods of extract emotional causes shown in Table 10.

5.3 Results on In-Context-Learning GPT

We observed similar trends in the case of the approach Zero-shot GPT-3.5 tends to only identify the neutral utterances accurately and failed in other categories. The incorporation of in-context learning improves the accuracy in identifying different emotion categories but there is little to no improvement in identifying disgust or fear utterances. The specific effect of emotional state extraction see Table 11.

6 Discussion

We compared all the methods and concluded that using large language models (LLMs) can improve the performance indicators of emotion recognition and cause analysis of the model, which is more effective than the traditional thinking chain based on the three-hop framework. The specific effects of emotional cause extraction on the test set are shown in Table 12. Some information can be obtained through the result indicators of the test data. For instance, Meta's closed-source GPT is often better than Llama's open-source large language model (LLM), and this has also been proven in this task. Due to the limitations of hardware conditions and research funds, we did not use the Llama model with larger parameters and the more expensive GPT-5 model. It is believed that better large models will lead to better results, and this will also be the future work of this task.

Table 8: Performance of multi-step THoR_cause with reasoning revision in speaker emotion cause extraction.

Approach	Epochs	Acc_state	F1_state	Acc_cause	F1_cause
THoR_cause with reasoning revision	0	58.859	44.756	78.128	42.663
	1	59.109	45.196	77.547	42.250
	2	**61.551**	**47.346**	**79.076**	**44.839**
	3	60.164	46.803	78.298	43.758

Table 9: Comparison of different Llama version model on speakers emotion state extraction.

Approach	Training Loss	Validation Loss	mode	ACC
Llama-2-7B	0.53	0.80	validation	0.66
Llama-2-13B	**0.47**	**0.73**	validation	**0.71**
Llama-3-8B	0.48	0.74	validation	0.67

7 Conclusion

The application of emotional cause analysis in the multimodal field is a relatively novel and highly challenging task because of many current mainstream large language models (LLMS) do not support the analysis of multimodal content data. In this task, the contents that need to be analyzed include text-based conversation records and those based on included audio-visual video clips that combine audio and visual modal data content. Although GPT-based method has made good use of these multimodal information and has also maximized the utilization of large language modalities to solve the problem with a certain type. However, there are still some areas that need improvement and deficiencies. The main shortcomings are analyzed below. Due to the insufficiency of personal computer computing power resources and the high cost of renting computing power, only the smaller parameter versions of the Llama series models can be used. This significantly affects the final prediction results of the multimodal emotional cause analysis and cannot fully utilize the best performance of the corresponding series of large language models (LLMs). When conducting emotional cause identification using GPT-3.5 and GPT-4V large language models (LLMs), it is not possible to perform local deployment and fine-tuning of the models. Instead, the online large language model (LLM) platform is utilized. Although during the prediction on test set data, demonstration examples and effective prompts were provided. However, there is still a lot of room for improvement. Future work will make further improvements based on these two aspects. As large language models (LLMs) continue to be updated and

Table 11: Emotion recognition results for seven emotion categories. P: precision, R: recall, F1: F1 score.

Approach	Metric	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Zero-Shot GPT	P	0.5652	0.2500	0.2727	0.4265	0.5385	0.5200	0.5906
	R	0.3333	0.4000	0.4286	0.5370	0.1842	0.3023	0.7426
	F1	0.4194	0.3077	0.3333	0.4754	0.2745	0.3824	0.6580
In-Context-Learning GPT	P	0.6667	0.2222	0.2222	0.4595	0.7000	0.5610	0.6957
	R	0.4615	0.4000	0.2857	0.6296	0.3684	0.5349	0.7059
	F1	0.5455	0.2857	0.2500	0.5312	0.4828	0.5476	0.7007

Table 12: The comparison of different method models for extracting the emotional causes of speakers has shown its effect on the test set.

Approach	w-avg. Strict F1	Strict F1
THoR-Cause-RR	0.3186	0.3243
Llama-2-7B	0.3117	0.3175
Llama-2-13B	0.3558	0.3630
Llama-3-8B	0.3259	0.3301
GPT-ICL	0.3982	0.4017

Algorithm 3 Pipeline of In-Context-Learning GPT method (GPT-based method)

Input: Video modality data, conversation data
Output: Emotion recognition, cause prediction
procedure VIDEOCAPTIONING(video)
    Generate conversation-level video caption using GPT-4V
    Output: Conversation-level video caption
end procedure
procedure EMOTIONRECOGNITION(caption)
    Retrieve training set conversation embeddings
    Perform vectorized similarity search
    Retrieve emotion-annotated conversation
    Add video caption content
    Use GPT-3.5 for emotion recognition
    Output: Emotion recognition results
end procedure
procedure CAUSEPREDICTION(caption)
    Retrieve training set context window embeddings
    Perform vectorized similarity search
    Retrieve cause-annotated context windows
    Add video caption content
    Use GPT-3.5 for cause prediction
    Output: Cause prediction results
end procedure
VIDEOCAPTIONING(video modality data)
EMOTIONRECOGNITION(conversation-level video caption)
CAUSEPREDICTION(conversation-level video caption)

Table 10: Comparison of different Llama version model on speakers emotion cause extraction.

Approach	Training Loss	Validation Loss	mode	Added pair
Llama-2-7B	0.51	0.83	validation	self-cause
Llama-2-13B	**0.45**	**0.75**	**validation**	**self-cause**
Llama-3-8B	0.48	0.81	validation	self-cause

Algorithm 2 Pipeline for fine-tuning Llama (Llama 2 and Llama 3)

Require: Conversational data $\mathcal{D}$, emotion categories, cause lists
Ensure: Fine-tuned LLMs for emotion recognition and cause prediction
procedure PROMPTGENERATION($\mathcal{D}$)
    for each conversation $c \in \mathcal{D}$ **do**
        Identify emotions in the utterances of each conversation c
        Generate prompts for emotion fine-tuning
        Format: `utterance_id :: emotion_label`
    end for
end procedure
procedure FINETUNELLM(prompts)
    Load pre-trained LLM (e.g., Meta Llama 2)
    Fine-tune LLM using prompts with Quantized Low-Rank Adaptation (QLoRA)
    Save fine-tuned model
end procedure
procedure GENERATEEMOTIONLABELEDCONVERSATIONS(test data, fine-tuned model)
    for each conversation c in test data **do**
        Predict emotions using fine-tuned model
        Generate emotion-labeled conversations
        Format: `utterance_id :: [emotion_label]`
    end for
    return emotion recognition results
end procedure
procedure FINETUNEANOTHERLLM(cause-labeled conversations)
    Load another pre-trained LLM (e.g., Meta Llama 2)
    Fine-tune model using cause-labeled conversations with Quantized Low-Rank Adaptation (QLoRA)
    Save fine-tuned model
end procedure
procedure INFERENCEONTESTSET(emotion-labeled test data, fine-tuned model)
    Predict causes on emotion-labled test set
    return cause prediction results
end procedure
prompts $\leftarrow$ PROMPTGENERATION($\mathcal{D}$)
$\text{model}_1 \leftarrow$ FINETUNELLM(prompts)
emotion-labeled test data $\leftarrow$ GENERATEEMOTIONLABELEDCONVERSATIONS(test data, model_1)
$\text{model}_2 \leftarrow$ FINETUNEANOTHERLLM(cause-labeled data)
INFERENCEONTESTSET(emotion-labeled test data, model_2)

iterated, the performance and capabilities of the models are constantly improving. The multimodal emotional cause task prediction based on large language models (LLMs) will also become more accurate.

References

[1] Soujanya Poria, Navonil Majumder, Rada Mihalcea, Eduard Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," IEEE Access (7), 100943–100953, 2019.

[2] Zixiang Ding, Huihui He, Mengran Zhang, Rui Xia, "From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification," In AAAI Conference on Artificial Intelligence (AAAI), 6343–6350, 2019.

[3] Fanfan Wang, Jianfei Yu, Rui Xia, "Generative emotion cause triplet extraction in conversations with commonsense knowledge," In Findings of the Association for Computational Linguistics: EMNLP, 3952–3963, 2023.

[4] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 527–536, 2019.

[5] Diman Ghazi, Diana Inkpen, Stan Szpakowicz, "Detecting emotion stimuli in emotion-bearing sentences," In International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 152–165, 2015.

[6] Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, Yu Zhou, "Event-driven emotion cause extraction with corpus construction," EMNLP, World Scientific, 1639–1649, 2016.

[7] Qinghong Gao, Jiannan Hu, Ruifeng Xu, Gui Lin, Yulan He, Qin Lu, Kam-Fai Wong, "Overview of ntcir-13 eca task," In Proceedings of the NTCIR-13 Conference, 2017.

[8] Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, Guodong Zhou, "An emotion cause corpus for chinese microblogs with multiple-user structures," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17(1), 1–19, 2017.

[9] Evgeny Kim, Roman Klinger, "Who feels what and why? annotation of a literature corpus with semantic roles of emotions" In Proceedings of the 27th International Conference on Computational Linguistics, 1345–1359, 2018.

[10] Laura Ana Maria Bostan, Evgeny Kim, Roman Klinger, "Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception," In Proceedings of The 12th Language Resources and Evaluation Conference, 1554–1566, 2020.

[11] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, "Recognizing emotion cause in conversations," Cognitive Computation, 1–16, 2021.

[12] Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, Erik Cambria, "Ecpec: emotion cause pair extraction in conversations," IEEE Transactions on Affective Computing, 2022.

[13] Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, Jianfei Yu, "Multimodal emotion-cause pair extraction in conversations," IEEE Transactions on Affective Computing, 14(3), 1832–1844, 2023.

[14] Fei, Hao, Li, Bobo, Liu, Qian, Bing, Lidong, Li, Fei, Chua, Tat-Seng, "Reasoning implicit sentiment with chain-of-thought prompting," Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 1171–1182, Association for Computational Linguistics, Toronto, Canada, 2023.

[15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, "Llama: Open and efficient foundation language models," 2023.

[16] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, "The dawn of lmms: Preliminary explorations with gpt-4(vision)," 2023.

[17] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, Zhifang Sui, "A survey for in-context learning," 2023.

[18] Rusnachenko, Nicolay, Liang, Huizhi, "nicolay-r at SemEval-2024 task 3: Using flan-t5 for reasoning emotion cause in conversations with chain-of-thought on emotion states," Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pp. 22–27, Association for Computational Linguistics, Mexico City, Mexico, 2024.

[19] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," Advances in Neural Information Processing Systems, 36, 2024.

[20] Chung, Hyung Won, Hou, Le, Longpre, Shayne, Zoph, Barret, Tay, Yi, Fedus, William, Li, Eric, Wang, Xuezhi, Dehghani, Mostafa, Brahma, Siddhartha, Webson, Albert, Gu, Shixiang Shane, Dai, Zhuyun, Suzgun, Mirac, Chen, Xinyun, Chowdhery, Aakanksha, Narang, Sharan, Mishra, Gaurav, Yu, Adams, Zhao, Vincent, Huang, Yanping, Dai, Andrew, Yu, Hongkun, Petrov, Slav, Chi, Ed H., Dean, Jeff, Devlin, Jacob, Roberts, Adam, Zhou, Denny, Le, Quoc V., Wei, Jason, "Scaling Instruction-Finetuned Language Models," arXiv, 2022.

[21] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, Lijuan Wang, "Mm-vid: Advancing video understanding with gpt-4(vision)," 2023.

[22] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, Diyi Yang, "Is chatgpt a general-purpose natural language processing task solver," 2023.