JMI at SemEval 2024 Task 3: Two-step approach for multimodal ECAC using in-context learning with GPT and instruction-tuned Llama models

Arefa[1,†], **Mohammed Abbas Ansari**[1,†], **Chandni Saxena**[2], **Tanvir Ahmad**[1]
[1]Jamia Millia Islamia University, New Delhi, India
[2]The Chinese University of Hong Kong, Hong Kong SAR, China
{arefa2001, mohd.abbas.ansari.2001}@gmail.com
csaxena@cse.cuhk.edu.hk, tahmad2@jmi.ac.in

Abstract

This paper presents our system development for SemEval-2024 Task 3: "The Competition of Multimodal Emotion Cause Analysis in Conversations". Effectively capturing emotions in human conversations requires integrating multiple modalities such as text, audio, and video. However, the complexities of these diverse modalities pose challenges for developing an efficient multimodal emotion cause analysis (ECA) system. Our proposed approach addresses these challenges by a two-step framework. We adopt two different approaches in our implementation. In Approach 1, we employ instruction-tuning with two separate Llama 2 models for emotion and cause prediction. In Approach 2, we use GPT-4V for conversation-level video description and employ in-context learning with annotated conversation using GPT 3.5. Our system wins rank 4, and system ablation experiments demonstrate that our proposed solutions achieve significant performance gains. All the experimental codes are available on Github.

1 Introduction

Emotion Cause Analysis (ECA) is centered around the extraction of potential cause clauses or pairs of emotion clauses and cause clauses from human communication, enabling a deeper understanding of communication dynamics. By incorporating multimodal cues like visual scenes, facial expressions, and vocal intonation, it facilitates a comprehensive and technically robust analysis of the factors that trigger diverse emotional reactions (Mittal et al., 2021; Zhang and Li, 2023; Zheng et al., 2023b). Despite the considerable amount of research conducted using diverse audio, visual, and text modalities (Gui et al., 2018; Xia and Ding, 2019; Fan et al., 2020; Shoumy et al., 2020; Abdullah et al., 2021), there has been a noticeable gap in the exploration of multimodal ECA in natural settings (human conversations). In this context, Wang et al. (2023a) introduce Multimodal Emotion Cause Analysis in Conversations (ECAC) task and provide Emotion-Cause-in-Friends (ECF) dataset, which incorporates text, audio, and video modalities. This task consists of two sub-tasks: Textual Emotion-Cause Pair Extraction in Conversations (Subtask 1) and Multimodal Emotion Cause Analysis in Conversations (Subtask 2). A detailed description of these sub-tasks can be found in the task description paper (Wang et al., 2024a).

In our submission to Subtask 2 of multimodal ECAC, this paper presents two distinct approaches to address the ECAC problem, giving competitive results. Drawing inspiration from the effectiveness of LLMs in diverse downstream tasks (Wang et al., 2023b, 2024b; Yang et al., 2024), including emotion recognition, we propose two LLM-based approaches that decompose the emotion-cause pair extraction process into two steps. The first step involves predicting the emotions of the utterances in the conversation. In the next step, we utilize these emotion labels to guide cause extraction. **Approach 1** involves instruction-tuning two separate Llama 2 models for emotion and cause prediction, while **Approach 2** leverages the in-context learning (ICL) capabilities (Dong et al., 2023) of the GPT-3.5 model. Additionally, we introduce an efficient technique using the GPT-4V model to extract conversation-level descriptions from video modality.

During the evaluation, our team ranked 4th on the leaderboard competing against more than 25 teams with a weighted-F1 score of 0.2816.

2 Background

2.1 Task definition

The input for the task, D, comprises N conversations. As described by Wang et al. (2023a), given a

[†]Equal contribution

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1561–1576

conversation $D_i = \{u_1, u_2, \ldots, u_M\}$ consisting of M utterances, where each utterance is represented by text, audio, and video, i.e. $u_j = [t_j, a_j, v_j]$, the goal of the task is to extract a set of emotion-cause pairs $P = \{\ldots, (u_k^e, u_k^c), \ldots\}$, where u_k^e denotes an emotion utterance and u_k^c corresponds to the cause utterance.

2.2 Related Work

The detailed Related Work section can be found in Appendix A.

2.3 Dataset

We use the Emotion-Cause-in-Friends (ECF) dataset provided by Wang et al. (2023a), which is summarized in Table 1. This dataset contains 13,509 multimodal utterances that occur in the American sitcom *Friends* with 9272 emotion-cause pairs. Each utterance consists of the text, video, and audio.

Class-distribution The dataset is imbalanced as shown in Fig. 1 wherein around 44% of the utterances have neutral emotion. Disgust and Fear constitute only 3% and 2.7% of the emotions.

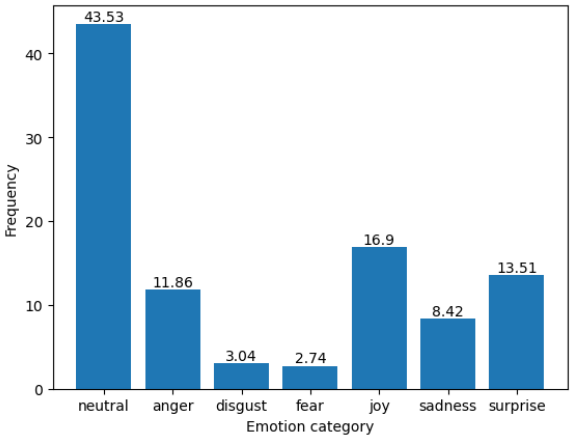

Figure 1: Percentage of each of the seven emotion categories

Relative positions of emotion and causes Interestingly, 49.95% of the causes are self-causes meaning that the same utterance caused itself as shown in Fig. 2. This is also intuitive, as what one speaks or expresses often elicits the emotion of their utterance. Note that the dataset curators have also annotated utterances coming after the emotion utterance as its cause. These constitute only about 2.8% of all causes and are one or two utterances away. 94.95% of the causes are 0-5 utterances behind the emotion utterance. The fact that what you speak or other interlocutors in the conversation speak affects the emotion of subsequent utterances explains this phenomenon.

Items	Number
Conversations	1344
Utterances	13,509
Emotional Utterances	7,690
Self-Causal Utterances	4,892
Non-Self-Causal Utterances	2,189
No Cause Emotional Utterances	609
Later-Causal Utterances	177

Table 1: Statistics of causes for emotional utterances.

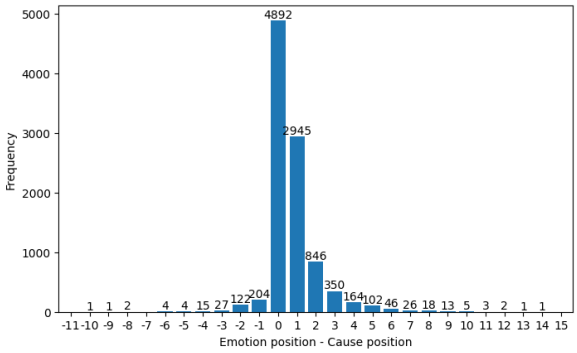

Figure 2: Relative position of emotion and causes

3 Methodology

3.1 Overview

We treat the task at hand as a two-step process. In the first step, we predict the emotion of each utterance in all N conversations. Here, the context C_j for utterance u_j of conversation D_i is the entire conversation itself. Given E target emotion labels and $\hat{y_j^e}$ as the predicted emotion label, the problem can be formulated as (where θ denotes the parameters):

$$\hat{y_j^e} = \arg\max_e \mathcal{P}(y^e | u_j, C_j, \theta) \quad (1)$$

In the second step, given these emotion labels, we predict the causes of each utterance that has an emotion other than neutral. The causes will be a subset of all utterances in the conversation D_i. Let the learned function be $f : U \rightarrow 2^U$, where U is the set of all utterances in the given conversation. It predicts the subset $\hat{y_j^c}$ of cause of emotion utterance u_j where $\hat{y_j^e} \neq$ neutral as:

$$\hat{y_j^c} = \arg\max_{y^c \in 2^U} \mathcal{P}(y^c | u_j, \hat{y_j^e}, C_j, \theta) \quad (2)$$

3.2 Approach 1: Fine-tuned Llama-2

In our first approach, we perform instruction fine-tuning of the Llama 2 Large Language Model, an open-source model developed by GenAI, Meta (Touvron et al., 2023). From the three variants with 7, 13, and 70 billion parameters, we use the 13 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 (Touvron et al., 2023). In addition, we use the Llama 2-chat version of the model[1], which is optimized for dialogue use cases as it aligns with our task. In our approach, we use Llama2 API[2] 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, given the prompt. Detailed explanations of these approaches are provided in the following sections. The fine-tuning procedure is shown in Fig.3.

3.2.1 Emotion recognition

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 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, as demonstrated by our ablation studies in Section 5. The instruction I_j^e, which gave the best results, is given in Appendix E.1 along with detailed prompt examples. The prompt consists of the instruction I_j^e and the context C_j for utterance u_j:

$$Prompt_j = (C_j, I_j^e) \tag{3}$$

Using this prompt as the input and the corresponding true emotion label y_j^e, we perform supervised fine-tuning of a Llama 2-13b model.

$$\hat{y_j^e} = \textbf{Llama}_e(Prompt_j, \theta) \tag{4}$$

We use a quantized version of the model due to memory limitations and perform Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2024) as a parameter-efficient fine-tuning technique. The training details are provided in the Section 4.

3.2.2 Cause prediction

To prepare the dataset for cause prediction, we incorporate the emotion labels obtained for each utterance. The conversation context now includes the emotion labels for each utterance u_j excluding those with a predicted emotion label $\hat{y_j^e}$ of *neutral*. This approach enhances the model's ability to analyze causal dependencies and identify which utterances may have contributed to a specific emotion. The output for cause prediction is a list of cause utterance IDs. The instruction is provided in Appendix E.1. The modified prompt for this step consists of this instruction I_j^c along with the conversational context with emotion labels C_j^e:

$$Prompt_j = (C_j^e, I_j^c) \tag{5}$$

Next, we perform supervised fine-tuning of a new Llama 2-13b model using this prompt as the input and the corresponding true list of causes:

$$\hat{y_j^c} = \textbf{Llama}_c(Prompt_j, \theta) \tag{6}$$

3.2.3 Adding video captions

To integrate cues from the videos corresponding to each utterance, we experimented using video captions generated using GPT-4 Vision as additional context for the model. However, we observed a notable decrease in performance since descriptions for individual utterances were somewhat noisy and did not effectively guide the predictions. Moreover, the captions often contained multiple emotions causing confusion for the model. As a result, we do not utilize these during training.

3.3 Approach 2: In-Context-Learning GPT

Our second approach (Fig. 4) tackles subtask 2 by obtaining conversation-level video captions using the GPT-4V(ision) model by OpenAI (Yang

[1]https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
[2]https://www.llama2.ai/

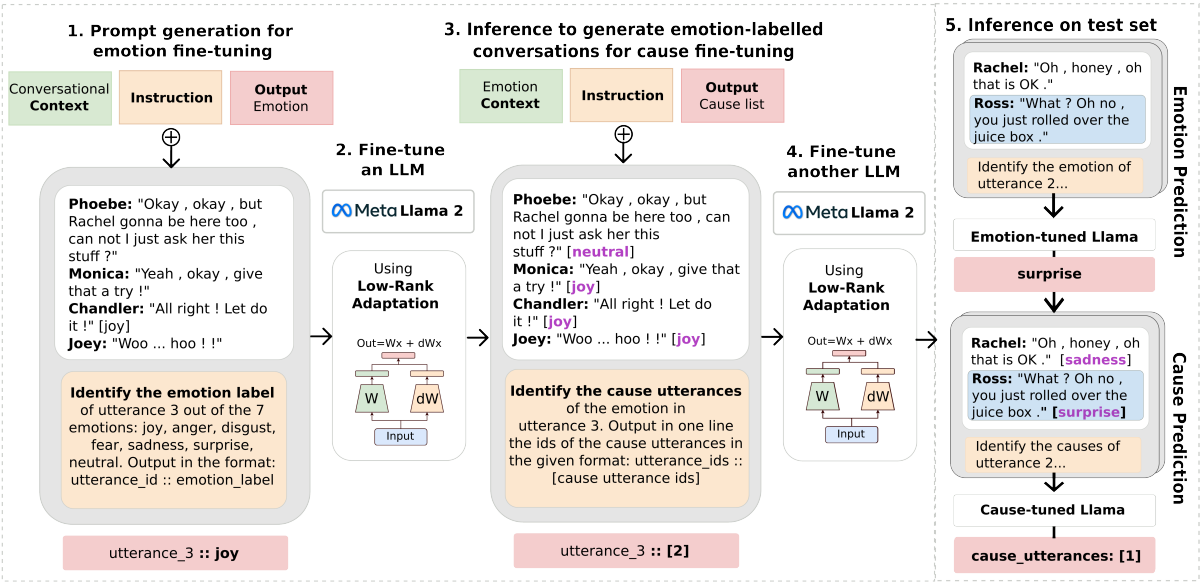

Figure 3: Pipeline for fine-tuning Llama (Approach 1)

et al., 2023). 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[3]. 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 (Dong et al., 2023) 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.

3.3.1 Video Captioning

GPT-4V has the capability to process video sequences (Yang et al., 2023; Lin et al., 2023). 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. The prompt is shown in the Fig. 5.

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. The process is illustrated in Fig. 5.

To accommodate the rate limits of the Vision API, we batch the utterances of a conversation and obtain outputs independently from the Vision model. We stitch all the outputs of a batched conversation into a single caption using GPT-3.5 (Appendix Fig. 16).

3.3.2 Emotion Recognition

GPT tends to be uncontrollable when performing zero-shot recognition of emotions in conversations (Qin et al., 2023) outputting emotions that are not a valid category of labels. 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 (Appendix Fig. 17). This retrieved conversation and its explanation serve as a demonstration for GPT to learn from, enabling it to recognize emotions in conversations more accurately. In addition, the prompt template includes the video caption as part of the input, as shown in Appendix Fig. 18.

To ensure effective ICL, it is important to provide general and descriptive examples that aid in solving the current task. In our approach, we sampled conversations from the training set containing all emotion categories. These conversations were stored as text-embedding-ada-002 embeddings (Neelakantan et al., 2022) in a vector database. At test time, we compute the embedding

[3]https://platform.openai.com/docs/models/gpt-3-5-turbo

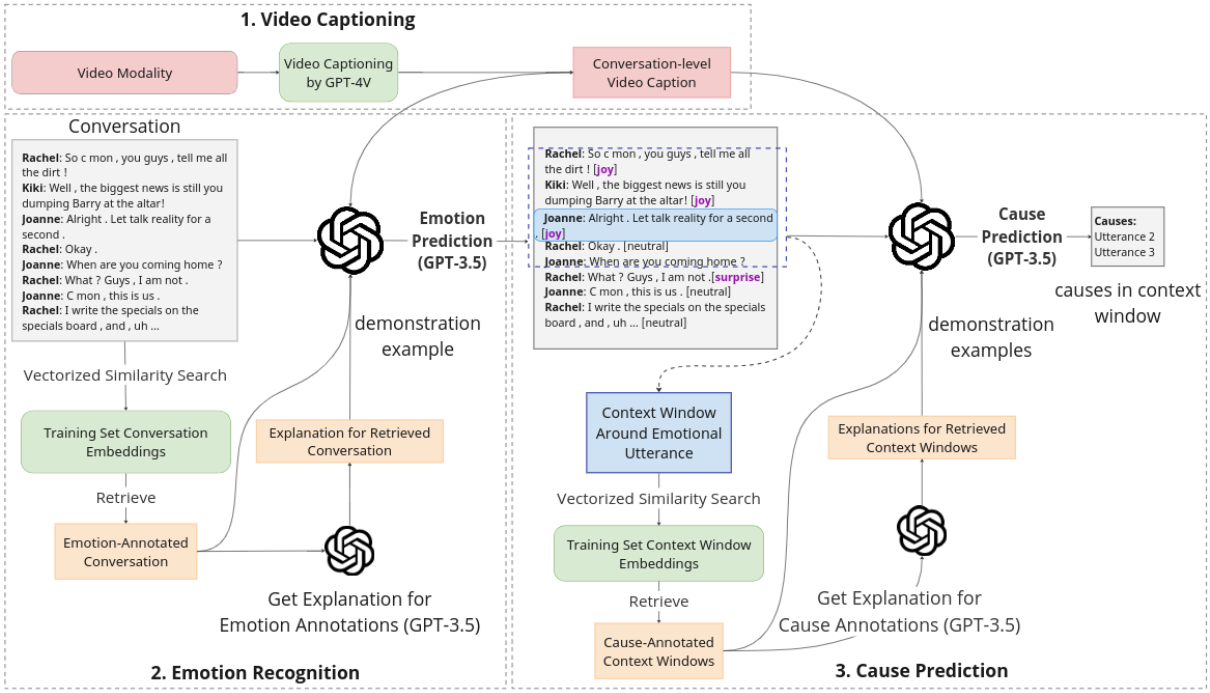

Figure 4: Pipeline of In-Context-Learning GPT Method (Approach 2)

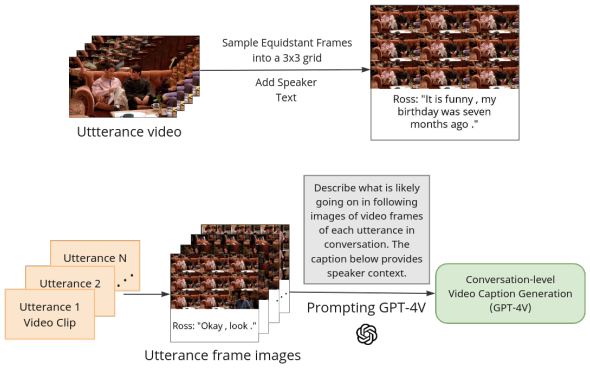

Figure 5: Video Captioning Pipeline

for a conversation and retrieve the closest matching embedding from the database based on Euclidean distance. The retrieved embedding aids ICL in improving emotion understanding and recognition.

3.3.3 Cause Prediction

Following the prediction of emotions, we predict the causes for each emotional utterance within a context window around that utterance. The bounds of the context window are given in Table 2. The bounds were informed by the distribution of the majority of relative positions of causes in the training set (Figure 2).

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 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 (Appendix Fig. 19). 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 (Appendix Fig. 20), since the local window may have lost some broader context.

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

Table 2: Context Window Bounds in each Direction

3.4 Post-Processing

In both our approaches, after getting the causes, we perform a post-processing step where we add the emotional utterance as its own cause which we call self-causes. This gives significant performance boosts as a majority of the causes are self-causes as pointed out in Appendix 2.3.

4 Experimental setup

Training details For approach 1, the data is split into train, test, and validation sets in the ratio 8:1:1.

We use peft library [4] for Parameter-Efficient Fine-Tuning. Due to memory constraints, we fine-tune a 4-bit quantized Llama-2 model using bitsandbytes library [5]. We report the details of the implementation for both approaches in Appendix B.

Evaluation metrics For evaluating, we report the precision, recall, F1-score, and weighted F1 which can be found on the competition website.[6]

5 Results and Discussion

Main results Both of our approaches gave competitive rankings on the official leaderboard for subtask 2 as shown in Table 3. In-context-learning GPT gave better results on the evaluation set compared to Fine-tuned Llama, thus our final position on the leaderboard was rank 4.

System	w-avg F1	F1
1. Samsung Research China-Beijing	0.3774	0.3870
2. NUS-Emo	0.3460	0.3517
3. SZTU-MIPS	0.3435	0.3434
4. GPT-ICL *(Ours)*	**0.2758**	**0.2816**
5. MotoMoto	0.2584	0.2595
6. Fine-tuned Llama *(Ours)*	**0.2558**	**0.2630**

Table 3: Leaderboard Results on Evaluation Data

Ablation study We conduct extensive ablation studies to measure the importance of the techniques we employ summarized in Table 4. For these experiments, we use a subset of our test set containing 528 utterances. It can be seen that the performance of zero-shot Llama as well as GPT is the lowest. Instruction-tuning and ICL clearly improve the performance on the task, showcasing the significance of making LLMs context-aware when tackling downstream tasks. Adding self-causes improves performance in both zero-shot and context-aware cases highlighting their importance. The incorporation of video captions leads to poorer results in context-learning. The detailed table is in Appendix C.

Limitations Our approaches are specific to one dataset and may not generalize well to other datasets. Due to resource limitations, we fine-tune a Llama 13b parameter model instead of 70b and use QLoRA instead of updating all parameters. To save costs, we used GPT-3.5 model instead of GPT-4. Even with extensive prompt engineering, GPT models tend to hallucinate or give unstructured outputs, requiring retry repeatedly.

Approach	F1	w-avg F1
Zero-shot Llama		
- w/o self-causes	0.117	0.116
- w/ self-causes	0.222	0.215
Instruction-tuned Llama		
- w/o self-causes	0.325	0.318
- w/ self-causes	**0.364**	**0.352**
Zero-shot GPT		
- w/o self-causes	0.100	0.097
- w/ self-causes	0.189	0.184
In-context-learning GPT		
- w/o self-causes w/o video	0.286	0.296
- w/o self-causes w/ video	0.235	0.241
-w/ self-causes w/o video	**0.336**	**0.342**
-w/ self-causes w/ video	0.329	0.334

Table 4: Results on Validation Set.

6 Conclusion

We tackled the Multimodal ECAC task with a two-step framework of recognizing emotions first and then predicting their causes using LLMs. We implemented two approaches: a Llama-2 model which has been fine-tuned with instructions and a GPT model which solves the task by learning from demonstration examples in context. Conversation-level video captions were extracted to provide more context to LLMs. Our second approach was our best submission for the task, placing us at rank 4 with our first approach being placed at rank 6. Our results were under cost constraints and further investigation with larger Llama-2 models and GPT-4 with more sophisticated ICL approaches are a clear follow-up of our work.

References

Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(02):52–58.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan

[4]https://huggingface.co/docs/peft/en/index
[5]https://github.com/TimDettmers/bitsandbytes
[6]https://nustm.github.io/SemEval-2024_ECAC/

Wermter. 2018. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Ying Chen, Wenjun Hou, Shoushan Li, Caicong Wu, and Xiaoqiang Zhang. 2020. End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 198–207.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187.

Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. Nnime: The nthu-ntua chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 292–298. IEEE.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *ArXiv*, abs/2301.00234.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. 2020. Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3707–3717.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th international conference on computational linguistics*, pages 4441–4453.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics*, 12(22):4714.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*, pages 152–165. Springer.

Lin Gui, Ruifeng Xu, Dongyin Wu, Qin Lu, and Yu Zhou. 2018. Event-driven emotion cause extraction with corpus construction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 145–160. World Scientific.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Guimin Hu, Guangming Lu, and Yi Zhao. 2021. Fss-gcn: A graph convolutional networks with fusion of semantic and structure for emotion cause analysis. *Knowledge-Based Systems*, 212:106584.

Mia Mohammad Imran, Preetha Chatterjee, and Kostadin Damevski. 2023. Uncovering the causes of emotions in software developer communication using zero-shot llms. *arXiv preprint arXiv:2312.09731*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework.

Weiyuan Li and Hua Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749.

Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Boundary detection with bert for span-level emotion cause analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 676–682.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, and Lijuan Wang. 2023. Mm-vid: Advancing video understanding with gpt-4v(ision). *ArXiv*, abs/2310.19773.

Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. 2021. Affect2mm: Affective analysis of multimedia content using emotion causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5661–5671.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al.

2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Sancheng Peng, Lihong Cao, Yongmei Zhou, Zhouhao Ouyang, Aimin Yang, Xinguang Li, Weijia Jia, and Shui Yu. 2022. A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*, 8(5):745–762.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2022. Deep emotion recognition in textual conversations: A survey. *arXiv preprint arXiv:2211.09172*.

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

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. 2005. Multimodal approaches for emotion recognition: a survey. In *Internet Imaging VI*, volume 5670, pages 56–67. SPIE.

Nusrat J Shoumy, Li-Minn Ang, Kah Phooi Seng, DM Motiur Rahaman, and Tanveer Zia. 2020. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, 149:102447.

Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. 2021. An end-to-end network for emotion-cause pair extraction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 84–91.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023a. Multimodal emotion-cause pair extraction in conversations. *IEEE Trans. Affect. Comput.*, 14(3):1832–1844.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024a. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2022–2033, Mexico City, Mexico. Association for Computational Linguistics.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024b. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023b. Augmenting black-box llms with medical textbooks for clinical question answering.

Yuwei Wang, Yuling Li, Kui Yu, and Yimin Hu. 2023c. Knowledge-enhanced hierarchical transformers for emotion-cause pair extraction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 112–123. Springer.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023d. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Jialiang Wu, Yi Shen, Ziheng Zhang, and Longjun Cai. 2024. Enhancing large language model with decomposed reasoning for emotion cause pair extraction. *arXiv preprint arXiv:2401.17716*.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

Shuntaro Yada, Kazushi Ikeda, Keiichiro Hoashi, and Kyo Kageura. 2017. A bootstrap method for automatic rule acquisition on emotion cause extraction. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 414–421. IEEE.

Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. 2024. Human-centric autonomous systems with llms for user command reasoning. In *Proceedings of the IEEE/CVF*

Winter Conference on Applications of Computer Vision (WACV) Workshops, pages 988–994.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *ArXiv*, abs/2309.17421.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. 2023. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, page 121692.

Xiaoheng Zhang and Yang Li. 2023. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13099–13110.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2024. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2023. Knowledge-bridged causal interaction network for causal emotion entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14020–14028.

Li Zheng, Donghong Ji, Fei Li, Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, and Chong Teng. 2023a. Ecqed: Emotion-cause quadruple extraction in dialogs. *arXiv preprint arXiv:2306.03969*.

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023b. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459.

Xiaopeng Zheng, Zhiyue Liu, Zizhen Zhang, Zhaoyang Wang, and Jiahai Wang. 2022. Ueca-prompt: Universal prompt for emotion cause analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7031–7041.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176.

A Related Work

Our system is designed to prioritize Subtask 2 which is directly related to text-based and multimodal ECA. In the following sections, we will present relevant research that addresses both unimodal (text-based) and multimodal ECA.

Text-based ECA

Advancements in text-based ECA (Xia and Ding, 2019; Hsu et al., 2018; Peng et al., 2022; Pereira et al., 2022) have made significant strides within the field of sentiment analysis. The task on emotion cause extraction (ECE) was initially proposed by Chen et al. (2010) on a Chinese corpus. Several studies (Li and Xu, 2014; Ghazi et al., 2015; Yada et al., 2017) have explored ECE task, using both rule-based and machine learning approaches that operate at the phrase or word level of the text data. Furthermore, (Gui et al., 2018) reformulated the ECE task as a clause-level classification problem and constructed a Chinese emotion-cause corpus based on the news data. Considering the effectiveness of clause-level units in indicating emotions, Xia and Ding (2019) introduced the task of Emotion-Cause Pair Extraction (ECPE) for extracting potential emotion-cause pairs from texts. Numerous deep learning models (Zhong et al., 2019; Wei et al., 2020; Chen et al., 2020; Singh et al., 2021; Li et al., 2021; Wang et al., 2023c) have been developed to address ECPE tasks. Additionally, graph-based approaches (Zheng et al., 2023a; Hu et al., 2021; Zhao et al., 2023) that utilize graphs to model dialog context and capture interactions between speakers and utterances hold significant potential. The focus on transformer models and the rapid progress in LLMs such as ChatGPT[7] and Llama (Touvron et al., 2023), have significantly boosted the performance of various NLP tasks (Imran et al., 2023) including ECPE (Wang et al., 2023d; Imran et al., 2023; Wu et al., 2024; Zheng et al., 2022).

[7] https://chat.openai.com/

Multimodal ECA

Given the strong association between facial cues and emotion, integrating modalities to improve emotion recognition has attracted a lot of attention (Sebe et al., 2005; Li et al., 2023; Zhang et al., 2023; Fu et al., 2023). Several key multimodal datasets (Wöllmer et al., 2013; Zadeh et al., 2016; Chou et al., 2017; Barros et al., 2018; Poria et al., 2019; Yu et al., 2020) have emerged to support and advance research. The availability of open conversation data has facilitated the expansion of multimodal conversation datasets, which includes various types of conversations such as dyadic interactions (Busso et al., 2008), and multi-participant communications (Hsu et al., 2018; Poria et al., 2019; Firdaus et al., 2020; Zheng et al., 2023b).

Large Language Models

The emergence of Large Language Models such as GPT-4 (Achiam et al., 2023), Llama (Touvron et al., 2023), PaLM (Anil et al., 2023), etc. has transformed the research landscape. Recently, there has been a surge in the application of LLMs to a multitude of domains. Zhang et al. (2024) extend their capabilities to the task of emotion recognition where they fine-tune a Llama 2-7 billion parameter model for emotion prediction. Lei et al. (2023) introduce a retrieval template module along with speaker identification and emotion-impact prediction tasks to improve the performance of LLM. In our work, as part of approach 1, we develop two distinct LLM-based experts separately for emotion and cause prediction.

Qin et al. (2023) investigated the task of zero-shot emotion cause prediction using ChatGPT with limited success. Recently, a new paradigm of in-context learning (ICL) (Dong et al., 2023) has emerged for LLMs that involves learning from a few examples to solve a variety of complex reasoning tasks (Wei et al., 2022b), (Wei et al., 2022a). Wu et al. (2024) proposed a Chain of Thought (CoT) (Wei et al., 2022b) approach for emotion cause pair extraction. Our approach 2 extends the idea of ICL towards solving the task of multimodal emotion cause pair extraction in two steps.

B Implementation details

B.1 Training details for Llama

Both emotion and cause prediction training used one Nvidia A100 40GB GPU for training (Available on Google Colab Pro priced at $11.8/month).

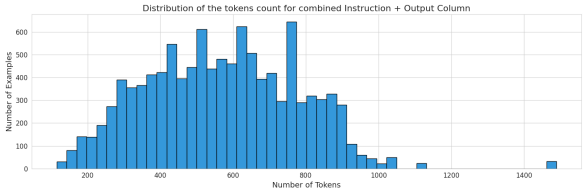

Figure 6: Distribution of token counts for Llama tokenizer

We train for one epoch due to constraints on Colab usage with gradient accumulation steps as 8 with an effective batch size of 8. A cosine learning rate scheduler and Adam optimizer are used. Inference is performed using two Tesla T4 16GB GPUs (Available on Kaggle for free (30 hrs/month)).

The long context length of 4096 tokens of the Llama 2 models, allows us to include the entire conversation as context and input that to the model. We perform experiments to analyze the maximum token counts in the dataset and observe that they do not exceed 1600 as shown in Figure 6. In case the token count exceeds the limit for the LLM we can use a window of utterances around the given utterance as context for predicting its emotion.

Hyperparameter	Value
Lora alpha	16
Lora dropout	0.1
Attention heads	16
Learning rate	1e-3
Epochs	1
LR scheduler	cosine
Warmup ratio	0.03
Weight decay	0.001

Table 5: Hyperparameters for fine-tuning

B.2 Details for in-context learning GPT

We use the LangChain[8] library to implement our three pipelines: video captioning, emotion recognition, and cause prediction. We use the interface provided by LangChain to communicate with OpenAI's API models detailed in Table 6.

Model	API Name
GPT-4V	gpt-4-vision-preview
GPT-3.5	gpt-3.5-turbo-1106
Embeddings	text-embedding-ada-002

Table 6: OpenAI API Model Names

Vector databases For creating vector databases, we use the FAISS Library (Douze et al., 2024). We

[8]https://github.com/langchain-ai/langchain

created a FAISS index containing embeddings of 12 conversations from the training set which contains all emotion categories. For cause prediction, we created a FAISS index for each of the 6 emotion categories and 3 possible positions of emotional utterance giving us a total of 18 indices. Each of these indices contained embeddings of context windows (bounds defined in Table 2) from the training set corresponding to each emotion and position.

C Detailed results

The detailed results on precision, recall, and F1-scores are given in Table 7.

D Error Analysis

We conduct error analysis for the output of emotion recognition using the two approaches. The performance of zero-shot Llama is extremely poor where the model predicts the label joy for almost all utterances (Fig. 7). On adding the conversational context, the model can identify the emotional nuances better, yet often predicts joy or surprise for neutral (Fig. 8). Instruction fine-tuning significantly boosts performance where the model can now differentiate distinct emotions (Fig. 9). 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 13, as shown in Table 8. We observed similar trends in the case of our second approach. Zero-shot GPT (Fig. 10) tends to only identify the neutral utterances accurately and fails in other categories. The incorporation of in-context learning (Fig. 11) improves the accuracy in identifying different emotion categories but there is little to no improvement in identifying disgust or anger utterances.

E Prompt details

E.1 Fine-tuned Llama 2

The general prompt for the Llama chat version is given in Figure 14. The prompts for emotion and cause prediction are given in Fig. 12 and Fig. 13. We provide a specific format for the output so as to ease the post-processing where we extract the first emotion label occurring after the "::" sequence of characters.

E.2 ICL-GPT

We devise prompt templates to be used in the LangChain framework. {} represent placeholders to be replaced when making a prompt. Video

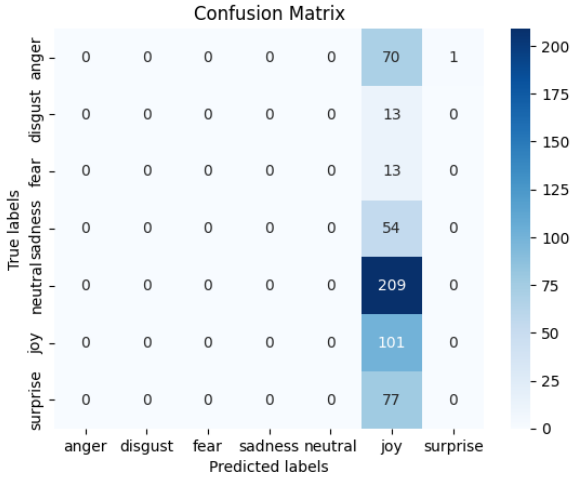

Figure 7: Confusion matrix for zero-shot emotion recognition without context using Llama

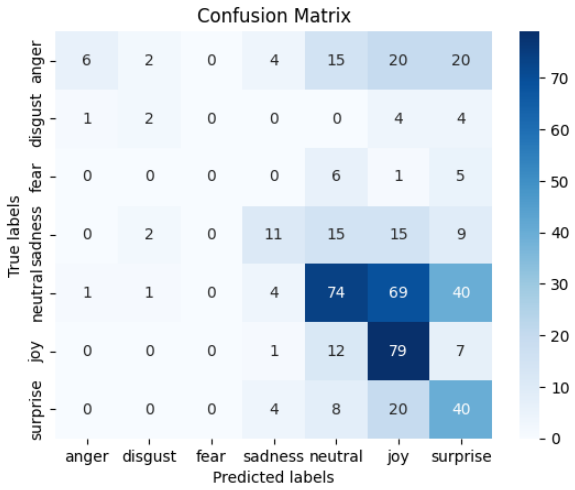

Figure 8: Confusion matrix for zero-shot emotion recognition with context using Llama

captioning prompt is given in Fig. 15. Due to rate limits, we had to batch the utterances, thus we may have multiple disjoint descriptions of a conversation. We prompt GPT-3.5 using the prompt in Fig. 16 to stitch the descriptions into a single caption. For explaining the retrieved conversation with emotion annotated, we use the prompt in Fig. 17. The retrieved conversation and explanation are now used as demonstration examples for the emotion recognition prompt in Fig. 18. For an explanation of causes in the retrieved-context window, we use the prompt in Fig. 19. The explanations of the retrieved windows are used as demonstration examples in the prompt for cause prediction within a context window as shown in the prompt in Fig. 20.

Approach	P	R	F1	w-P	w-R	w-avg F1
Zero-shot Llama w/o self-causes	0.089	0.168	0.117	0.090	0.168	0.116
Zero-shot Llama w/ self-causes	0.157	0.372	0.222	0.152	0.372	0.215
Instruction-tuned Llama w/o self-causes	0.351	0.304	0.325	0.335	0.304	0.318
Instruction-tuned Llama w/ self-causes	**0.360**	**0.367**	**0.364**	**0.342**	**0.367**	**0.352**
Zero-shot GPT w/o self-causes	0.081	0.130	0.100	0.087	0.130	0.097
Zero-shot GPT w/ self-causes	0.140	0.290	0.189	0.149	0.290	0.184
In-context-learning GPT w/o video captions w/o self-causes	0.259	0.319	0.286	0.283	0.319	0.296
In-context-learning GPT w/o video captions w/ self-causes	**0.270**	**0.445**	**0.336**	**0.287**	**0.445**	**0.342**
In-context-learning GPT w/o self-causes	0.216	0.256	0.235	0.241	0.256	0.241
In-context-learning GPT w/ self-causes	0.261	0.445	0.329	0.280	0.445	0.334

Table 7: Results on Validation Set. P: precision, R: recall, w: weighted.

Approach	Metric Supp	Anger 71	Disgust 13	Fear 13	Joy 101	Sadness 54	Surprise 77	Neutral 209
Zero-shot Llama w/o context	**P**	0.0000	0.0000	0.0000	0.1881	0.0000	0.0000	0.0000
	R	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
	F1	0.0000	0.0000	0.0000	0.3166	0.0000	0.0000	0.0000
Zero-shot Llama with context	**P**	0.7500	0.2857	0.0000	0.3798	0.4583	0.3200	0.5663
	R	0.0845	0.1538	0.0000	0.7822	0.2037	0.5195	0.4498
	F1	0.1519	**0.2000**	0.0000	0.5113	0.2821	0.3960	0.5013
Fine-tuned Llama with context	**P**	0.5641	0.0	0.3333	0.6210	0.625	0.6103	0.6666
	R	0.6197	0.0	0.1538	0.5842	0.3704	0.6104	0.7943
	F1	**0.5906**	0.0	**0.2105**	**0.6020**	**0.4651**	**0.6104**	**0.7249**
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 8: Emotion Recognition Results for Seven Emotion Categories. P: precision, R: recall, F1: F1 score, and Supp: support.

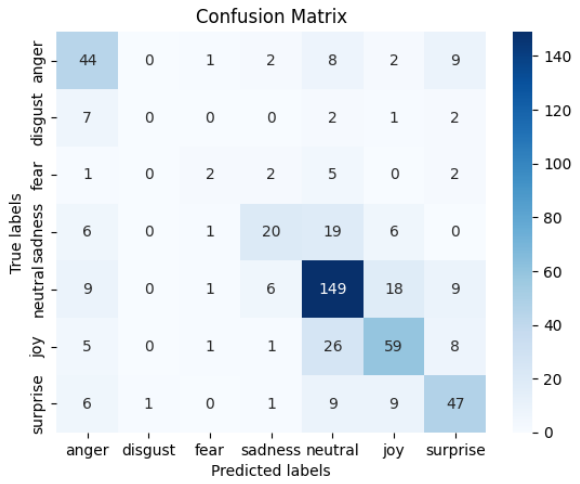

Figure 9: Confusion matrix for emotion recognition with context using fine-tuned Llama

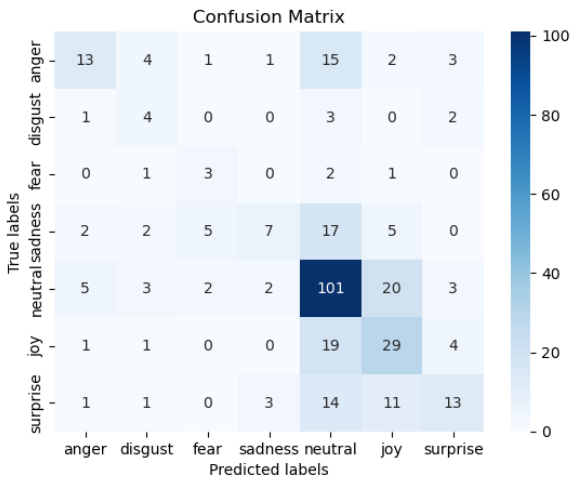

Figure 10: Confusion matrix for emotion recognition using Zero-shot GPT-3.5

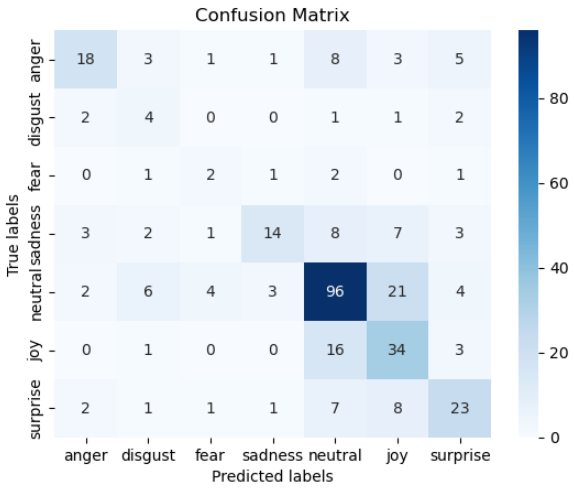

Figure 11: Confusion matrix for emotion recognition using GPT-ICL

Emotion Recognition Prompt:

<s>[INST]
"conversation": [
{
"utterance_ID": 1,
"text": "This is just Bactine . It will not hurt .",
"speaker": "Monica"
},
{
"utterance_ID": 2,
"text": "Sorry , that was wax .",
"speaker": "Joey"
},
{
"utterance_ID": 3,
"text": "Oh , poor little Tooty is scared to death . We should find his owner .",
"speaker": "Phoebe"
},
{
"utterance_ID": 4,
"text": "Why do not we just put poor little Tooty out in the hall ?",
"speaker": "Ross"
},
{
"utterance_ID": 5,
"text": "During a blackout ? He would get trampled !",
"speaker": "Rachel"
}]
Identify the emotion label of utterance 5 out of the 7 emotions: anger, fear, disgust, sadness, joy, surprise, neutral. Don't give an explanation.
Output only one line in the format:
utterance_id :: emotion_label
[/INST]
utterance_5 :: anger
</s>

Figure 12: Example Prompt for emotion prediction using Llama

Cause Prediction Prompt:
<s>[INST]
"conversation": [
{
"utterance_ID": 1,
"text": "This is just Bactine . It will not hurt .",
"speaker": "Monica",
"emotion": "neutral"
},
{
"utterance_ID": 2,
"text": "Sorry , that was wax .",
"speaker": "Joey",
"emotion": "neutral"
},
{
"utterance_ID": 3,
"text": "Oh , poor little Tooty is scared to death . We should find his owner .",
"speaker": "Phoebe",
"emotion": "sadness"
},
{
"utterance_ID": 4,
"text": "Why do not we just put poor little Tooty out in the hall ?",
"speaker": "Ross",
"emotion": "disgust"
},
{
"utterance_ID": 5,
"text": "During a blackout ? He would get trampled !",
"speaker": "Rachel",
"emotion": "anger"
}]
Identify the cause utterances of the emotion in utterance 5. Output in one line the ids of the cause utterances as a list in the given format:
utterance_id :: [cause utterance ids]
Don't give any explanation.
[/INST]
utterance_id :: [4,5]
</s>

Figure 13: Example Prompt for cause prediction using Llama

General Prompt Template:
\<s>[INST] <<SYS>>
{{system message}}

<</SYS>>

{{message/input}}
[/INST]
{{answer}}
</s>

Figure 14: General Prompt Template for Llama

Video Caption Prompt:
You are an expert of Friends TV Show. You can understand a video scene from a few of its frames shown in sequence. You give precise descriptive analysis.
Describe what is likely going on in following images of video frames of each utterance in conversation. The caption below provides speaker context.
Give output as:
Scene Description:

Figure 15: Video Captioning Prompt Template

Caption Stitching Prompt:
Following is a descriptions of video clip from Friends TV show for a particular conversation. The descriptions are broken from each other. Stitch the description into a continous coherent narrative of the whole scene

Figure 16: Batched Video Caption Stitching Prompt Template

Emotion Explanation Prompt:
There are 6 basic emotions: Anger, Disgust, Fear, Joy, Sadness, Surprise.
The emotion of the speaker is determined by the context of the conversation.
If the emotion is not in any category, is a mix of several categories, or is ambiguous it can be categorized as "Neutral".
Analyze the following conversation where emotion of each utterance is annotated in square brackets at the end.
Give reasoning behind the annotation of each utterance.

{conversation}

Output a JSON in the following format:
[{{"utterance_ID": id,
"text" : content,
"speaker": speaker
"emotion": emotion,
"explanation": detailed explanation}}
...
]
No plain text.

Figure 17: Emotion Label Explanation Prompt Template

Emotion Recognition Prompt:
You are a die-hard fan of the popular Friends TV show.
You have all the knowledge of all the seasons and are familiar with all the characters.
Your task is to recognize emotions in utterances.
Here's an annotated example with recognized emotion and explanation:
{example}

Like above example annotate the following Conversation:
Context for the scene is given below:
{scene}

Conversation:
{conversation}

Classify the emotional state of the speaker in each utterance into ONLY one out of the 6 emotions:
Anger, Disgust, Fear, Joy, Sadness, Surprise.
The emotion of the speaker is determined by the context of the conversation.
Give explanation for your classification using the context. Only Use the above 6 emotion categories.
If the emotion is not in any category, is a mix of several categories, or is ambiguous,
classify the state as "Neutral". Sarcastic comments may be categorized as Neutral.
Format the output as JSON as the given example. No plain text.

Figure 18: Emotion Recognition with Context Learning Prompt Template

Cause Explanation Prompt:
You are an expert in analyzing conversations to extract the causes of emotions
in particular utterances by speakers. You give definite confident answers only.
Description of emotional causes:
- Each utterance always has a reason of why it was said and why it had a particular emotion.
- A cause is an utterance that comes before or after the particular utterance in question that best explains to be the reason behind the particular emotion.
- The emotional utterance itself can be a cause of itself if its content ALSO best explains the reason for the particular emotion.
- Sometimes the cause can be beyond the context of the conversation thus an utterance might have no cause within conversation
- There can be multiple causes for an utterance.

Here's a conversation:
{conversation}

Analyze and justify the above annotation concisely.

Figure 19: Cause Explanations Prompt Template

Cause Prediction Prompt:
You are an expert in analyzing conversations to extract the causes of emotions
in particular utterances by speakers. You give definite confident answers only.
Description of emotional causes:
- Each utterance always has a reason of why it was said and why it had a particular emotion.
- A cause is an utterance that comes before or after the particular utterance in question that best explains to be the reason behind the particular emotion.
- The emotional utterance itself can be a cause of itself if its content ALSO best explains the reason for the particular emotion.
- Sometimes the cause can be beyond the context of the conversation thus an utterance might have no cause within conversation
- There can be multiple causes for an utterance.

Here are some examples of how to recgonize causes:
Example 1:
{example_1}

Example 2:
{example_2}

Example 3:
{example_3}

Now, please recognize the causes in following conversation. Heres the context for the whole conversation:
{scene}

Conversation:
{window}

Figure 20: Cause Prediction with Context Learning Prompt Template