




Bootstrap Wayfinding Questions to Elicit Emotion Shift Reasoning with Large Language Models

Vy Nguyen , Xiuzhen Zhang* , Feng Xia 

School of Computing Technologies, RMIT University, Victoria, Australia
s3964796@student.rmit.edu.au, xiuzhen.zhang@rmit.edu.au, feng.xia@rmit.edu.au

Abstract—Emotions are integral to communication, and understanding emotions is critically important for healthcare applications such as mental well-being and empathetic support. However, in conversational exchanges, emotions constantly shift as conversations unfold. Existing studies on reasoning about emotion shift triggers (ESR) require extensive manual annotation of both utterance-level emotion labels and emotion shift triggers. In this research, we propose a novel approach, the Wayfinding Instruction Tuning (WIT) framework, to tackle the ESR challenge in a more efficient manner with minimal dependency on emotion labels. WIT employs a large language model (LLM) to generate sequences of dynamic, context-sensitive wayfinding questions designed to guide the instruction tuning of another LLM on the ESR task. In contrast to conventional prompting, our key innovation is a structured *wayfinding prompting strategy*, which systematically decomposes the ESR process into carefully designed sub-questions. Together with their ground-truth-aligned answers, these sub-questions guide the model to learn how to detect changes in emotion and identify their triggers. Our experiments show that, even when fine-tuned on a medium-sized dataset of 100 training instances, using only pre- and post-shift emotion labels, WIT can directly predict emotion shifts without first identifying the emotion of each utterance. This design avoids intermediate emotion classification and its associated error propagation, a common issue in two-step pipelines. WIT achieves comparable performance (F1: 0.749 vs 0.760) to state-of-the-art supervised models while requiring only shift-boundary labels during training and zero annotation at inference, providing a practical alternative when complete utterance labels are unavailable.

Index Terms—Emotion Analysis, Large Language Models, Natural Language Processing

I. INTRODUCTION

EMOTIONS are an integral part of communication. Early emotion studies have predominantly focused on detecting and classifying the emotional content of the text. In communication, however, one’s emotions always change dynamically as a conversation progresses. Understanding the triggers behind one’s emotional dynamics provides explanations for the underlying drivers of their decision making [1]–[6].

Figure 1 presents a conversation exhibiting three shifts in the emotions of the speakers. Utterances 3 *triggers* Tom’s first emotion shift from anger to surprise, then utterance 6 *triggers* both Tom and Lily to change their emotions. These shifts and their triggers create a dynamic interplay of empathy, understanding, and emotional regulation in their conversation.

In current literature on emotion shift reasoning (ESR), most prior studies [2], [7]–[12] on identifying emotion shift triggers

rely on human annotations of emotion labels for each utterance to train models, which requires substantial manual annotation efforts and is costly. Another class of studies formulates ESR as a multi-task learning problem [13], including an emotion recognition (ER) sub-task. But errors in the ER stage can propagate, leading to a cascading impact on the downstream ESR task.

Earlier work on emotion shift detection [14] focused on identifying whether shifts occurred rather than their triggering utterances. Recent studies [15]–[27] leverage LLMs for emotion analysis in conversations using textual or multi-modal data, but primarily address emotion recognition rather than trigger identification. Beyond emotion analysis, advances in LLM reasoning have shown promise for complex tasks. Prompting strategies like Chain-of-Thought (CoT) [28] and its extensions enable models to generate intermediate reasoning steps [29]–[32], while instruction tuning has been shown to significantly improve LLMs’ contextual understanding and adaptability [33]–[37] for various tasks. Recent research has further demonstrated that LLMs possess substantial emotional intelligence capabilities [38]–[40], providing empirical support for leveraging LLMs in emotion shift reasoning tasks.

In this paper, we aim to develop a more efficient approach to ESR that reduces the reliance on explicit emotion annotations. Towards this goal, we need to address the challenges of how to identify emotion shifts and their triggers. The complexity arises from emotions’ subtle, context-dependent nature [41]–[43]. Conversations include implicit cues, sarcasm, and cultural nuances [30], [44], [45] requiring contextual understanding of conversational settings, relationships, and beliefs [46], [47].

To address these challenges, we propose a novel framework called **Wayfinding Instruction Tuning (WIT)**. WIT begins by prompting an LLM θ_1 to dynamically generate **wayfinding instructions**, each of which is a sequence of ground-truth-aligned question-answer pairs designed to capture the nuances of emotion shift triggers. These instructions are then used to tune another LLM θ_2 , teaching it to learn this reasoning process. Once tuned, θ_2 can be applied to predict for unlabelled conversations. This time, the wayfinding questions are generated in the same manner but are presented sequentially to the tuned θ_2 , which answers them step by step.

By constructing this reasoning chain, WIT navigates the complexity of conversational dynamics and detects emotion triggers effectively. For the conversation in Figure 1, θ_1 generates a series of questions, starting with auxiliary ones

*Corresponding author

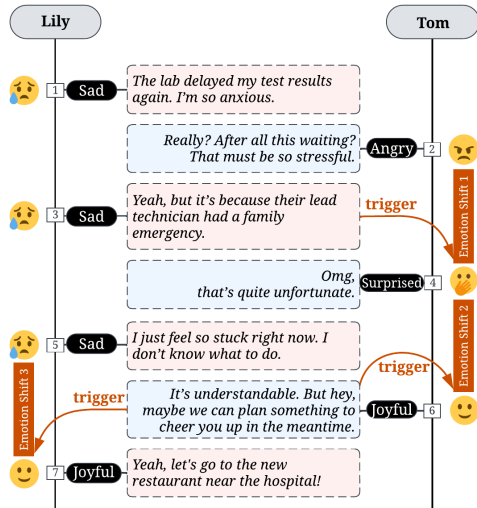


Fig. 1: A sample conversation with three emotion shifts, each of which is triggered by a preceding or current utterance.

that guide the thought process and culminating in a final task-specific question. These questions are then presented sequentially to the tuned θ_2 , which provides corresponding answers, as illustrated in Figure 2. The process concludes when the emotion triggers are identified.

This step-by-step decomposition of the conversation is reminiscent of the process suggested by cognitive appraisal theories of emotion [48]–[50], where humans evaluate events in multiple stages (e.g., assessing significance, potential outcomes, coping resources) before arriving at an emotional shift. By breaking down the context and eliciting sub-questions about changing perspectives and triggers, WIT aligns with how humans systematically interpret emotional dynamics in real-world dialogues.

To our knowledge, this is the first study on ESR requiring only shift-boundary labels rather than complete utterance-level annotations. Our WIT framework enhances reasoning and adaptability of LLMs for ESR by introducing:

- *Structured reasoning for ESR*: We propose a method for LLMs to explore emotion dynamics in conversations using emotion-specific heuristics, effectively tracing sequential emotion shifts. Building upon CoT, we decompose conversation analysis into a sequence of questions, guiding the model to track emotional transitions and pinpoint triggering utterances by directly predicting emotion shifts, without first labeling the emotion of each utterance.
- *Diverse reasoning instructions*: We utilise multiple LLMs to generate varied phrasings or sequencing of wayfinding questions. This approach is especially beneficial when training data are limited, as it creates multiple complementary reasoning paths from the same instance, enriching the instruction-tuning data.
- *Sequential instruction tuning for emotion shift reasoning*: We fine-tune LLMs using task-specific wayfinding questions that encode ESR-specific reasoning: tracking emotional trajectories, encoding pre- and post-shift emotion

Direct Prompt, Pretrained LLM

Q: Locate any emotion shifts and identify the exact utterances that trigger them.

A: Tom's emotion shift occurs when Lily explains that the lead technician's family emergency. At first, he focuses on validating her feelings, showing empathy and understanding toward her frustration about the repeated delays. This initial response highlights his support for her emotional state and acknowledgment of the stress she is experiencing. However, when Lily provides additional context about the technician's situation, Tom's focus shifts. Instead of continuing to align with her frustration, he transitions to expressing concern and sympathy for the technician, recognising the gravity of their personal challenges. This shift demonstrates Tom's ability to adapt his emotional response based on new information, moving from shared frustration to a more compassionate perspective.

Wayfinding Questions, Tuned LLM

Q: What emotion does Tom initially express in response to Lily's concern?

A: Tom initially expresses anger in response to Lily's concern.

Q: How does Lily's clarification about the delay add new information to the conversation?

A: It explains the delay was due to a family emergency of the lab technician.

Q: How might this new information challenge or soften Tom's initial emotional response?

A: It reframes the situation as unfortunate rather than purely frustrating.

Q: Identify the position of the utterances that trigger Tom's emotional shift.

A: The *third utterance* in the conversation triggers Tom's emotion shift.

Fig. 2: Typical LLM reasoning (left) and a series of wayfinding questions that identifies the utterance that triggers Tom's emotion to shift in Figure 1 (right).

states, and identifying triggering utterances.

Experiments on widely used conversational datasets show that our WIT framework achieves performance comparable to state-of-the-art supervised models while requiring substantially less annotation effort.

II. RELATED WORK

This section reviews existing research on ESR and related LLM technologies.

A. Emotion Shift Reasoning

Research on emotion shift detection [14] has evolved from identifying whether shifts occurred to serving as an auxiliary task for improving Emotion Recognition (ER) models. Gao et al. [51] used a Bi-LSTM-based emotion shift detection (ESD) module to generate shift representations, which were then concatenated with emotional semantic representations. Wang et al. [13] proposed MtlERC-ES, a multi-task learning model simultaneously performing ER, ESD, and sentiment classification. Tu et al. [52] introduced a persona-infused cross-task graph network, which first captured speakers' relationships via a persona-infused refinement network, then used cross-task connections for ESD as a precursor to ER.

Identifying triggers for emotion shifts in conversations was explored by Kumar et al. [2], who introduced two approaches: a multi-task Transformer-based network without utterance-level emotion labels that performed sub-optimally, and a single-task network trained with gold emotion annotations that performed significantly better. Later, Kumar et al. [7], [53] expanded this work by examining factors like shock or threat as instigators of emotion shifts. More recently, Nguyen and Zhang [8] proposed a supervised framework using stacked instruction fine-tuning of LLMs to directly address ESR. Similarly, Shaik et al. [54] combined instruction tuning with quantised low-rank adaptation techniques. Tian et al. [9] presented XGBoost-EC, a DL approach identifying emotion

provocateurs, while Wan et al. [10] used two GRU-based models to tackle ESR. Unlike prior work requiring complete utterance annotations, WIT requires only shift-boundary labels (pre/post-shift emotions), substantially reducing annotation overhead while maintaining competitive performance.

B. LLM Prompting

Prompting guides LLMs by framing the input in a manner that encourages desired responses [55]. Amongst the most popular strategies is CoT reasoning, a technique that entails generating intermediate reasoning steps before providing the conclusion [28], thus enabling LLMs to tackle complex problems requiring reasoning [29]. Building on CoT, Yao et al. [56] and Long [57] introduced Tree-of-Thoughts, aiming to use multifold chained intermediate reasoning steps to encourage models to explore a broader range of thought processes. Similarly, Zhang et al. [58] extended CoT to multimodal contexts by integrating both textual and visual modalities, resulting in more comprehensive rationales.

For more complex tasks, it is crucial to design dynamic prompts tailored to the specific nature of the data. Zhang et al. [59] introduced PromptGen, a pioneering framework to dynamically generate prompts for knowledge probing, leveraging pre-trained generative models. Shum et al. [60] proposed AutomateCoT, a strategy that automates CoT prompting by generating, pruning, and selecting rationale chains from a small labelled dataset. Similarly, Akella et al. [61] developed an auto-prompt generation system for multiple LLMs with minimal training, combining reinforcement learning for task-relevant sequencing and cell-level similarity for enhancing example selection. Expanding on CoT, Dorbala et al. [62] introduced wayfinding CoT, which uses LLMs to construct dynamic reasoning steps for robot navigation tasks.

Our approach is motivated by Dorbala et al. [62], but the key contribution lies in the task-specific design of wayfinding questions tailored to the emotion shift reasoning (ESR) task, rather than the use of sequential instruction tuning itself. Unlike spatial waypoints, our questions explicitly model emotional dynamics by encoding pre- and post-shift emotion states, tracking speaker-level emotional trajectories, and identifying triggering utterances from implicit conversational cues. Their sequential structure reflects the staged nature of emotion shift reasoning [48], [49], where each step builds on prior emotional context to progressively narrow down the trigger. This task-specific reasoning design is supported by our ablation study, which shows that progressively adding ESR-oriented questions consistently improves F1 scores across all models.

C. Instruction Tuning

Instruction tuning, which involves tuning LLMs on datasets of $(Input; Response)$ pairs in a supervised way, equips them with a deeper understanding of how to interpret and respond to task-oriented requests [15]. In research on emotions, Luo et al. [63] developed an emotion-cause-aware instruction-tuning approach to fine-tune multiple language models, aiming to enhance their ability to recognise emotions and their causes.

Similarly, for the same task, Etal et al. [16] performed instruction tuning with two separate LLAMA 2 models, one for recognising emotions and the other for predicting their causes. Building on this, Cheng et al. [17] introduced Emotion-LLaMA, a multimodal model that integrates audio, visual, and textual inputs. Through instruction tuning, Emotion-LLaMA demonstrated significant improvements in emotion recognition (ER) performance. In summary, existing studies on leveraging LLMs focus on emotion recognition rather than emotion shift reasoning, and their approaches are based on single instruction-tuning.

More recently, Hu et al. [64] proposed sequential fine-tuning LLMs for complex problems, where a sequence of instructions is used to enhance LLM reasoning to solve complex problems. They showed that LLMs fine-tuned using a multi-step approach (e.g., “translate, then answer” or “describe, then answer”) outperformed those tuned on direct instructions like “answer the question.” While instruction tuning is effective for enhancing task-oriented responses, sequential instruction tuning serves as the training mechanism for our ESR-specific wayfinding questions. Our WIT framework falls into this line of research, and importantly, for reasoning about the complex emotion dynamics, we develop strategies to prompt an LLM to automatically generate wayfinding questions (and answers) as pseudo instructions to fine-tune another LLM.

III. METHODOLOGY

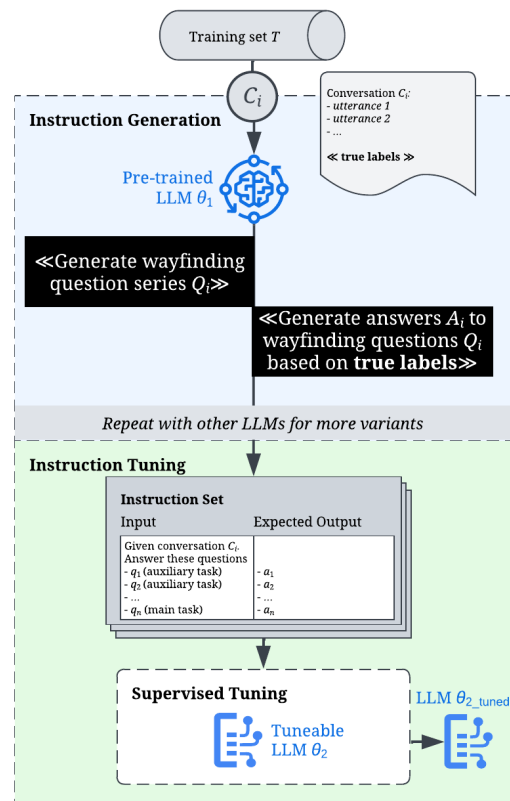


Fig. 3: The architecture of the WIT framework. Two steps are involved: instruction generation with the first LLM and instruction tuning of the second LLM.

In this section, we propose the WIT framework to address the ESR problem. This problem can be formulated as follows: Given a conversation between multiple speakers involving n utterances, presented in the format $C = \{(s_1, u_1), (s_2, u_2), \dots, (s_n, u_n)\}$. Any speaker s_i can appear again in a different utterance. For all pairs of utterances u_x and u_y that express the emotions e_x and e_y respectively, where $x < y$, $s_x = s_y$, and $e_x \neq e_y$, identify a set of utterances $T \in \{u_1, u_2, \dots, u_y\}$ that trigger speaker s_x (or s_y) to shift their emotion from e_x to e_y .

The WIT framework is a supervised instruction-tuning framework in which an LLM θ_1 generates a *wayfinding* instruction set to help tune another LLM θ_2 to identify emotion shift triggers. It consists of two steps: *instruction generation* and *instruction tuning*. The architecture of this framework is presented in Figure 3.

A. Instruction Generation

At the heart of our approach is the *wayfinding prompting strategy*, which extends standard CoT methods by specifying dimension-specific alignment with the ESR task, completeness of emotional cues, contextual relevance, logical sequencing of queries, and clarity of expression. The WIT framework uses this to prompt a pre-trained LLM θ_1 to generate a series of wayfinding questions that facilitate the identification of emotion shift triggers.

For a conversation C_i in the training set T , these wayfinding questions constitute a series $Q_i = \{q_1, q_2, \dots, q_k\}$ that emulate the stepwise human cognitive process for understanding emotional transitions. Our prompt template for prompting θ_1 to generate Q_i is as follows:

Task: You are tasked with generating a series of questions to guide the reasoning process for identifying {...}'s emotion shifts and their triggers in a conversation.

Input: Below is a conversation between multiple speakers:
{...}

Each utterance STRICTLY expresses only one of the following emotions: {anger, contempt, disgust, fear, joy, sadness, surprise, neutral}.

Output Requirements: Generate a series of wayfinding questions that adhere to the following criteria:

- *Logical Progression:* Each question should build upon the information revealed in the answer to the previous question.
- *Context Awareness:* The questions must explicitly focus on the emotional context of the speaker and the conversational flow.
- *Clarity:* Ensure each question is precise and easily interpretable.
- *Completeness:* Cover all aspects of the reasoning process, from initial emotion detection to the identification of emotion shifts and their triggers.
- *Final Task-Specific Question:* Conclude with a question that asks for the exact position(s) of the trigger utterance(s).

Example Output Format: Provide the questions as a series without explanation:

- Question 1: [Initial question]
- Question 2: [Follow-up question based on Question 1]
- ...
- Final Question: [A question asking to specify the EXACT position(s) of the trigger utterance(s)].

Generated questions focus on shift detection, not utterance-level classification. This design requires only shift-boundary

labels during training (e.g., *neutral* in utterance 3 \rightarrow *joy* in utterance 7) and zero annotations during inference, while supervised methods require utterance-level emotion annotations during training and predict per-utterance emotions at inference before identifying shifts, introducing error propagation.

Subsequently, we use only pre- and post-shift emotion labels from the training set to re-prompt θ_1 and generate answers to each wayfinding question. We validate instruction quality through (1) automated checks ensuring trigger positions \leq conversation length and emotions within valid sets; (2) manual spot-checking of 50 randomly sampled instances by the authors, verifying that predicted trigger positions match ground-truth labels, showing 96% accuracy; (3) multi-LLM generation reducing individual model biases. The prompt template is as follows:

Task: You are tasked with generating relevant answers to a provided series of questions based on the truths provided.

Input: Below is a conversation between multiple speakers:
{...}

You are provided with the following question series:
{...}

Output Requirements: Using the truths provided below, generate concise and accurate answers to each question in the series.

Truths Provided:

- *Emotion Shift:* {...} shifts from {...} in utterance {...} to {...} in utterance {...}.
- *Trigger Utterance(s):* {...}.

Instructions:

- Align your answers to the reasoning implied by the questions.
- Maintain logical consistency and clarity in the progression of answers.
- Avoid commentary or explanations outside the context of the provided truths.

These answers, $A_i = \{a_1, a_2, \dots, a_k\}$, are intrinsically grounded in the conversation's content, thereby closely approximating a natural reasoning process. This process results in a sequential question-answer series for the conversation, represented as $\{q_i \rightarrow a_i\}_{i=1}^k$, in which each answer is explicitly tied to the ground-truth labels. This explicit alignment allows us to construct a granular, human-like reasoning process.

B. Instruction Tuning

Using the wayfinding question series generated, we proceed to instruction-tune another LLM θ_2 . An instruction t_i comprises two parts: the *input* and the *desired output*. In the WIT framework, the input contains the conversation C_i and the series of questions Q_i , whereas the desired output contains the series of answers A_i to the questions. t_i can be presented as follows:

$$t_i = \ll \langle \{C_i, Q_i\}; A_i \rangle \gg$$

The questions Q_i represent a sequence of tasks, comprising multiple auxiliary tasks and a final main task. The auxiliary tasks guide the model through the context, aiming to gather the necessary information and insights required to effectively perform the main task.

Instruction t_i needs to be tokenised as $t_i = \{t_i(1), t_i(2), \dots, t_i(n)\}$, where $t_i(j)$ denotes the j -th

token of t_i and n is the number of tokens, before it can be used for the training. This token sequence can be divided into two distinct parts by the index l where the content of the instruction input ends, namely the instruction input part $t_i (< l)$ and the desired response part $t_i (\geq l)$. Now, given the tokenised instruction dataset $\mathcal{S} = \{t_i\}_{i=1}^N$, where N is the size of the set, supervised instruction tuning can be performed using cross-entropy loss as follows [65]:

$$\mathcal{L} = \sum_{t_i \in \mathcal{S}} \mathcal{L}_i$$

$$\mathcal{L}_i = - \sum_{j=1}^{|t_i|} \log P(t_i(j) | t_i(< j); \theta_2)$$

In this function, for each token $t_i(j)$, the model θ_2 predicts the next token given all preceding tokens $t_i(< j)$. Loss is computed only on the output portion ($j \geq l$), training the model to generate correct answers given conversation context and questions. This sequential multi-task tuning—where auxiliary questions guide the model toward the final trigger identification task—is expected to enhance reasoning capabilities more effectively than direct single-task tuning. Upon completion, we obtain a tuned model $\theta_{2_{\text{tuned}}}$.

For inference, the initial θ_1 is used to generate wayfinding questions from unlabelled data. The generated questions are then prompted to $\theta_{2_{\text{tuned}}}$, which now produces the corresponding answers. These answers are then processed programmatically to extract the relevant labels, thereby completing the inference process.

IV. EXPERIMENTS & RESULTS

Our experiments focus on Ekman’s seven discrete emotions: *anger*, *contempt*, *disgust*, *fear*, *joy*, *sadness*, and *surprise* [66]. We also use *neutral* to label utterances that do not express an explicit emotion.

A. Experiment Setup

Datasets: To ensure generality, we employ three datasets for evaluation: MELD-FR [2], EFR-MASAC [7], and ConvECPE [5]. Statistics regarding these datasets are presented in Table I. MELD-FR comprises English dialogues sourced from the popular American TV series *Friends* [67]. EFR-MASAC is derived from Hindi-English code-mixed video clips of the popular Indian series *Sarabhai vs. Sarabhai* [68]. ConvECPE, in particular, was developed as a conversational corpus aimed at extracting emotion-cause pairs [5]. For this study, we adapted the ConvECPE dataset to suit the ESR experiment by converting the existing cause labels into trigger labels.

Baselines:

Supervised baselines using utterance labels: This class trains emotion classifiers from complete utterance-level annotations then predicts shifts. Comparing against these methods tests whether WIT achieves comparable performance without full utterance labelling. Baselines include: (1) Stacked instruction tuning [8], where LLMs are tuned with instructions built on emotion labels (F1: 0.76 MELD-FR, 0.79 EFR-MASAC);

| Dataset | Split | Instances | Triggers | Utterances | % Trigger |
|-----------|-------|-----------|----------|------------|-----------|
| MELD-FR | Train | 4,000 | 5,575 | 35,000 | 15.9% |
| | Val | 426 | 494 | 3,522 | 14.0% |
| | Test | 1,002 | 1,169 | 8,642 | 13.5% |
| EFR-MASAC | Train | 4,894 | 6,542 | 98,777 | 6.6% |
| | Val | 389 | 434 | 7,462 | 5.8% |
| | Test | 385 | 461 | 7,690 | 6.0% |
| ConvECPE | Train | 120 | 1,249 | 5,810 | 21.5% |
| | Test | 31 | 393 | 1,623 | 24.2% |

TABLE I: Statistics for the MELD-FR, EFR-MASAC, and ConvECPE datasets.

(2) XGBoost-EC [9] using emotion provocateurs; (3) Two-GRU model [10] with BERT embeddings; (4) Rule-based algorithms [12].

Multi-task baseline: EFR-ERC_{multi} [2] operates without utterance-level labels but still predicts per-utterance emotions before shifts, enabling comparison of direct shift detection versus two-stage approaches.

In-context learning: Zero-shot and one-shot prompting with GPT-4o, LLAMA 3 8B, and Mistral 7B assess LLM capabilities without task-specific tuning.

Framework settings: For each training dataset, an LLM (GPT-4o is used in our experiments) is first employed to select top m instances based on two criteria: (1) the selected subset must maintain a similar distribution with the original training data in terms of the number of speakers, utterances, and triggers; and (2) the chosen dialogues should exhibit diversity, avoiding redundancy or excessive similarity.

Instruction Generation with θ_1 We generate $m \times t$ series of wayfinding questions and answers, where m ($= \max, 500, 100, 20$) denotes the top labelled instances selected and t denotes the number of variants generated for each (Table II). These values of m represent large, medium, and small training set sizes. Distribution is preserved. Variants are generated using different LLMs to enrich the training data. For MELD-FR and EFR-MASAC, 6 different instruction sets are generated. For ConvECPE, only 5 instruction sets are generated, because the dataset size is not large enough for the *Many instances* test.

For test scenarios involving a single series, we use GPT-4o to generate it. When three series variants are required for the same instance, we utilise GPT-4o, a candidate of frontier models and two open-source mid-sized models, LLAMA 3 8B and Mistral 7B, to produce them.

| Test Scenario | m | t |
|--|-----|-----|
| Full training data (WIT _{max} ×1) | max | 1 |
| Many instances (WIT ₅₀₀ ×1) | 500 | 1 |
| Medium instances (WIT ₁₀₀ ×1) | 100 | 1 |
| Medium instances, 3 variants (WIT ₁₀₀ ×3) | 100 | 3 |
| Limited instances (WIT ₂₀ ×1) | 20 | 1 |
| Limited instances, 3 variants (WIT ₂₀ ×3) | 20 | 3 |

TABLE II: Test suite design for instruction generation.

An advantage of this strategy is harnessing multiple LLMs $\{\theta_{1_1}, \theta_{1_2}, \dots\}$ to generate diverse question sequences from single instances. Each LLM produces unique phrasings and

| Method | MELD-FR | EFR-MASAC | ConvECPE |
|---|--------------|--------------|--------------|
| Baselines using emotion labels for supervision | | | |
| Stacked Instruction _{+labels} [8] | 0.760 | 0.790 | - |
| XGBoost-EC _{+labels} [9] | 0.710 | 0.790 | - |
| Two GRUs _{+labels} [10] | 0.680 | 0.790 | - |
| Algorithms _{+labels} [12] | 0.680 | 0.790 | - |
| Baselines using Multi-task learning | | | |
| EFR-ERC _{multi} [2] | 0.348 | - | - |
| Baselines using in-context learning | | | |
| GPT-4o + Zero-shot | 0.353 | 0.253 | 0.378 |
| GPT-4o + One-shot | 0.334 | 0.310 | 0.335 |
| LLAMA + Zero-shot | 0.300 | 0.261 | 0.360 |
| LLAMA + One-shot | 0.365 | 0.307 | 0.294 |
| Mistral + Zero-shot | 0.348 | 0.281 | 0.335 |
| Mistral + One-shot | 0.325 | 0.271 | 0.344 |
| WIT tuned models | | | |
| GPT-4o + WIT _{max×1} | 0.743 | 0.783 | 0.729 |
| GPT-4o + WIT _{500×1} | 0.749 | 0.704 | - |
| GPT-4o + WIT _{100×1} | 0.670 | 0.680 | 0.721 |
| GPT-4o + WIT _{100×3} | 0.694 | 0.736 | 0.750 |
| GPT-4o + WIT _{20×1} | 0.594 | 0.448 | 0.523 |
| GPT-4o + WIT _{20×3} | 0.629 | 0.510 | 0.571 |
| LLAMA + WIT _{max×1} | 0.734 | 0.656 | 0.736 |
| LLAMA + WIT _{500×1} | 0.709 | 0.637 | - |
| LLAMA + WIT _{100×1} | 0.729 | 0.665 | 0.731 |
| LLAMA + WIT _{100×3} | 0.739 | 0.703 | 0.746 |
| LLAMA + WIT _{20×1} | 0.584 | 0.375 | 0.590 |
| LLAMA + WIT _{20×3} | 0.619 | 0.482 | 0.645 |
| Mistral + WIT _{max×1} | 0.687 | 0.600 | 0.731 |
| Mistral + WIT _{500×1} | 0.643 | 0.696 | - |
| Mistral + WIT _{100×1} | 0.673 | 0.685 | 0.717 |
| Mistral + WIT _{100×3} | 0.695 | 0.738 | 0.740 |
| Mistral + WIT _{20×1} | 0.466 | 0.402 | 0.547 |
| Mistral + WIT _{20×3} | 0.503 | 0.435 | 0.548 |

TABLE III: Performance of models across all datasets. Values are F1 scores for the triggers. Models denoted with _{+labels} are those requiring pre-existing utterance-level emotion labels. Bold values denote the best performance in each local group. Highlighted values represent our best fine-tuned model for each dataset. Multiple methods report identical performance; values verified from original papers.

reasoning paths, enabling diverse training data from limited examples (see Appendix ?? for examples).

Instruction Tuning with θ_2 We tune three LLMs, GPT-4o, LLAMA 3 8B, and Mistral 7B on all the instruction sets generated for each dataset. As a result, we have 66 different tuning pipelines. We utilise the AdamW optimiser with a learning rate of $1e^{-5}$ and a cosine annealing scheduler. Batch sizes are dynamically adjusted based on GPU memory availability. The tuning is conducted over 3 epochs to balance performance and time while avoiding overfitting on the limited dataset sizes.

B. Main Results

Our primary evaluation metric is the F1 score of the emotion triggers, labelled as 1. This choice is justified by the central objective of identifying trigger utterances for emotion shifts.

WIT achieves performance comparable to supervised models with reduced annotation. WIT models achieve F1 scores within 0.011 (MELD-FR) and 0.007 (EFR-MASAC) of supervised SOTA requiring full utterance labels (Table III),

demonstrating competitive performance with substantially reduced supervision. Specifically, GPT-4o+WIT_{500×1} achieves 0.749 versus 0.760 for supervised baselines on MELD-FR, while GPT-4o+WIT_{max×1} reaches 0.783 versus 0.790 on EFR-MASAC.

Optimal configuration with moderate training data. The WIT_{100×3} configuration proves optimal, delivering 7 of 9 best results across datasets. Performance plateaus beyond 100 instances with careful instruction curation, though fewer than 20 instances shows significant degradation. For MELD-FR, GPT-4o achieves best performance (F1: 0.749) with WIT_{500×1}, while LLAMA and Mistral excel with WIT_{100×3} (F1: 0.739, 0.695). Similar patterns hold for EFR-MASAC and ConvECPE, with WIT_{100×3} consistently delivering strong results. Open-source LLMs achieve performance comparable to GPT-4o, enabling cost-effective implementation.

Annotation efficiency. During training, WIT requires only shift-boundary labels (e.g., 4 labels for a 20-utterance dialogue with 2 shifts) versus complete utterance-level labels (20 labels) for supervised methods. More importantly, during inference, WIT directly detects emotion shifts and their triggers with zero annotation, while supervised methods must first predict per-utterance emotions then identify shifts, introducing computational overhead and error propagation.

C. Evaluation of Wayfinding Questions

We evaluate wayfinding question quality using Gemini-1.5-Pro-002, scoring *alignment*, *completeness*, *relevance*, *sequencing*, and *clarity* (1-10 scale) for WIT_{100×3} (Table IV). Two human annotators evaluated three representative conversations: C1, a casual conversation; C2, a monologue with no emotional changes; and C3, a multi-speaker dialogue with three triggers (Appendix ??).

| | GPT-4o | LLAMA 3 8B | Mistral 7B |
|--------------|--------|------------|------------|
| Alignment | 7.64 | 6.49 | 6.87 |
| Completeness | 8.12 | 6.75 | 6.46 |
| Relevance | 8.53 | 7.54 | 7.21 |
| Sequencing | 8.26 | 6.43 | 6.75 |
| Clarity | 8.93 | 8.18 | 8.34 |

TABLE IV: Average scores by Gemini for WIT_{100×3} instruction set.

Our evaluation reveals notable differences across models. The automated evaluation using Gemini-1.5 (Table IV) further confirms model-level differences in instruction quality. GPT-4o outperforms the open-source models across all dimensions, particularly in clarity (8.93) and relevance (8.53), indicating well-structured, contextually grounded question generation. LLAMA and Mistral show comparable performance in clarity but lag in alignment, sequencing, and completeness, suggesting less coherent task decomposition and weaker progression in reasoning chains. These results align with human evaluations and reinforce that high-quality wayfinding prompts benefit from both linguistic fluency and strategic reasoning scaffolding, which GPT-4o executes more reliably.

Qualitative evaluations (Table V) reveal that GPT-4o consistently delivers high-quality questions across all criteria, with

particularly strong sequencing and completeness. For example, in C3, “*What is the tone of Ross’s statements (lines 1–7)?*” effectively guides attention across utterance progression. In contrast, LLAMA 3 8B and Mistral 7B show weaker alignment and sequencing, with questions like “*Does Monica’s emotion shift occur after a statement [...]?*” that lack contextual grounding and fail to reflect cumulative emotional buildup, instead isolating single utterances rather than modeling broader discourse context.

| Criteria | GPT-4o | | | LLAMA 3 8B | | | Mistral 7B | | |
|--------------|--------|----|----|------------|----|----|------------|-----|----|
| | C1 | C2 | C3 | C1 | C2 | C3 | C1 | C2 | C3 |
| Alignment | 8.5 | 9 | 9 | 6 | 6 | 6 | 6 | 6 | 7 |
| Completeness | 8 | 8 | 9 | 6.5 | 5 | 7 | 7 | 7 | 7 |
| Relevance | 8 | 8 | 8 | 7 | 7 | 7 | 6.5 | 6.5 | 6 |
| Sequencing | 9 | 8 | 9 | 7 | 5 | 5 | 7 | 7 | 7 |
| Clarity | 9 | 7 | 8 | 7.5 | 8 | 7 | 8 | 8 | 7 |

TABLE V: Annotators’ evaluation of the sample conversations. Values are average scores. Inter-annotators reliability: Cohen’s $\kappa = 0.68$ (substantial agreement).

D. Error Analysis

Our error analysis focuses on the top-performing variant of base model, namely GPT-4o + WIT_{500×1}, LLAMA + WIT_{100×3}, and Mistral + WIT_{100×3}, to evaluate their ability to identify emotional triggers. Figure 4 presents the confusion matrices these models produce. GPT-4o demonstrates a balanced performance, with a 5.4% false positive rate and an 8.6% false negative rate, making it the most versatile for general tasks. LLAMA achieves exceptional precision for non-trigger utterances, with only 0.5% false positives, but struggles with trigger detection, exhibiting a high false negative rate of 41.2%, limiting its suitability for trigger-sensitive applications. Mistral balances low false positives (2.5%) with a moderate false negative rate (16.25%), offering a compromise but underperforming in trigger detection compared to GPT-4o. While each model’s strengths are evident, the elevated false negatives in LLAMA and Mistral highlight areas for improvement in accurately detecting emotional triggers.

| Predictions | True Labels | | True Labels | | True Labels | |
|-------------|------------------------------|--------------|------------------------------|--------------|--------------------------------|--------------|
| | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 7403 85.66% | 70 0.81% | 7190 93.50% | 190 2.47% | 1199 73.88% | 144 8.87% |
| 1 | 427 4.94% | 742 8.59% | 39 0.51% | 271 3.52% | 427 4.94% | 742 8.59% |
| | GPT-4 + WIT _{500×1} | | LLAMA + WIT _{100×3} | | Mistral + WIT _{100×3} | |

Fig. 4: Confusion matrices for top performing instruction-tuned GPT-4o, LLAMA 3 8B, and Mistral 7B models. 0 denotes non-trigger utterances whereas 1 denotes trigger utterances.

E. Evaluation of Emotion Shift Predictions

Table VI shows that GPT-4o + WIT_{500×1} achieves high accuracy on frequent, high-contrast shifts—e.g., *neutral* → *joy*

(198 total, 142 correct; 71.7%) and *joy* → *neutral* (179/139; 77.7%). These transitions involve clear polarity changes and strong lexical cues, which are well-aligned with WIT’s step-wise question decomposition. In contrast, the model performs poorly on rare or fine-grained shifts like *contempt* → *fear* and *fear* → *disgust*, where emotion boundaries are more ambiguous and contextual cues subtle. These errors reflect limitations in generalising over low-frequency transitions and capturing latent interpersonal or cultural nuance.

Shifts between negatively-valenced emotions (e.g., *anger* → *sadness*, *disgust* → *contempt*) yield intermediate accuracy, indicating that while the model detects affective change, it struggles to precisely differentiate among negative emotional states. Notably, transitions involving *surprise* (e.g., *surprise* → *anger*, 41/31) are predicted with reasonable accuracy, suggesting the model handles emotionally dynamic or reactive shifts better than gradual ones. These results affirm WIT’s strength in modelling dominant emotional arcs but highlight its challenges in low-resource or semantically adjacent transitions.

| Before \ After | Anger | Cntmpt | Disgust | Fear | Joy | Neutral | Sadness | Surprise |
|----------------|-------------------|------------------|------------------|------------------|--------------------|--------------------|------------------|------------------|
| Anger | | 19 ₁₀ | 17 ₁₁ | 12 ₈ | 29 ₂₀ | 101 ₇₈ | 19 ₁₄ | 41 ₃₁ |
| Contempt | 5 ₃ | | 1 ₀ | 2 ₀ | 12 ₁₉ | 25 ₁₅ | 1 ₀ | 2 ₁ |
| Disgust | 12 ₉ | 1 ₁ | | 2 ₁ | 4 ₃ | 25 ₁₉ | 5 ₃ | 7 ₅ |
| Fear | 5 ₂ | 1 ₀ | 1 ₀ | | 7 ₅ | 37 ₂₈ | 7 ₄ | 5 ₃ |
| Joy | 18 ₁₄ | 3 ₁ | 3 ₂ | 6 ₃ | | 179 ₁₃₉ | 29 ₂₂ | 39 ₂₈ |
| Neutral | 102 ₇₅ | 25 ₁₅ | 19 ₁₄ | 34 ₂₅ | 198 ₁₄₂ | | 57 ₄₇ | 85 ₆₅ |
| Sadness | 26 ₂₁ | 2 ₁ | 2 ₁ | 5 ₃ | 26 ₂₀ | 60 ₄₂ | | 18 ₉ |
| Surprise | 34 ₂₃ | 4 ₃ | 8 ₆ | 3 ₂ | 30 ₂₀ | 99 ₇₉ | 14 ₁₁ | |

TABLE VI: GPT-4o + WIT_{500×1} emotion shift predictions across all datasets. Cell values denote the frequency of an emotion shift, while subscripts denote accurate predictions.

To further investigate the errors, we examine a case study of a code-mixed Hindi-English conversation taken from the EFR-MASAC dataset involving three speakers, Monisha, Indu, and Sahil. Appendix ?? provides the dialogue, the wayfinding questions generated by GPT-4o, and the answers to them generated by the LLAMA + WIT_{100×3} model. The dialogue is humorous and dynamic, featuring playful banter, overlapping speaker interactions, and frequent language switches. Monisha, intending to deliver a speech about her mother-in-law, ends up presenting a comically off-topic piece on milk due to plagiarism from a primary school essay collection. Sahil and Indu mock her, further adding to the complex interplay of emotions and misunderstandings. The tuned LLAMA model identifies Monisha’s emotion shift from *surprise* to *fear* in utterance 9, but ground truth labels show a shift from *contempt* in utterance 10 to *neutral* in her final response. The emotional trigger arises from Sahil’s challenge in utterance 15 and Monisha’s self-explanation in utterance 16.

The errors likely arise from the linguistic ambiguity of the code-mixed dialogue, overlapping speaker interactions, and the comedic tone, which masks deeper emotions like contempt. This example underscores the difficulty of recognising emotional shifts in conversations that are linguistically mixed, involve multiple speakers, and hinge on subtle self-reflective triggers in comedic and layered contexts.

F. Ablation Analysis

We assess sequential tuning efficacy by decomposing the tuning pipeline using the WIT_{100×1} instruction set for MELD-FR. Instead of the full series, we start by tuning models with only the final question, then progressively add preceding questions. This reveals the impact of sequential instruction tuning.

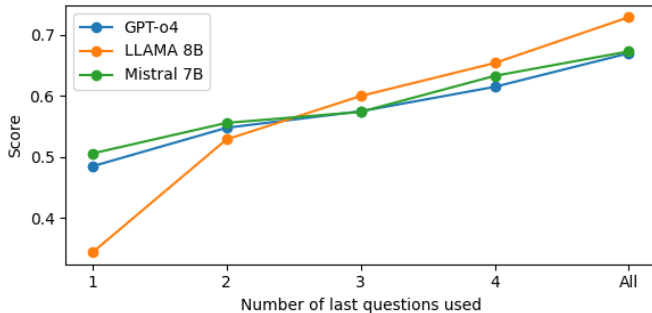


Fig. 5: Ablation study results for models instruction-tuned on WIT_{100×1} instruction set.

Progressive question addition. Figure 5 shows that F1 scores gradually increase as more guiding questions are added. When tuned with only the final question, models achieve suboptimal F1 scores of 0.485 (GPT-4o), 0.344 (LLAMA), and 0.506 (Mistral). As questions are added, models progressively align their reasoning patterns, reaching 0.670, 0.729, and 0.673 respectively with the full instruction set.

Component ablations. To isolate which components drive performance, we conducted ablations on GPT-4o with WIT_{100×1} for MELD-FR (baseline F1: 0.670). Removing individual requirements from the wayfinding prompt template degrades performance: without Logical Progression (F1: 0.609), without Context Awareness (F1: 0.590), and minimal prompt with no requirements (F1: 0.569). Furthermore, altering question sequence also reduces effectiveness: random ordering (F1: 0.630), reversed ordering (F1: 0.611), and final-question-only without auxiliary questions (F1: 0.485). These results confirm that both structured progression through carefully ordered questions and task-specific prompt requirements are essential for effective ESR.

G. Hallucinations

The appearance of emotion labels beyond Ekman’s set (e.g., *frustration, embarrassment*) is common in model outputs. We do not consider these as hallucinations, as they represent valid emotional experiences not captured by Ekman’s taxonomy. However, hallucinations occur when models predict triggers beyond conversation length or reference non-existent utterances (see Appendix ??). To address this, we implement automated validation checking: (1) trigger positions \leq conversation length, (2) emotion labels within valid set. Invalid outputs are rejected and regenerated (mean attempts: 1.2, max: 3, success rate: 98.7%). While this re-run protocol introduces minor selection bias, the low attempt rate suggests negligible impact.

| Model | MELD-FR | EFR-MASAC | ConvECPE |
|------------|---------|-----------|----------|
| GPT-4o | 48 | 32 | 19 |
| LLAMA 3 8B | 59 | 75 | 28 |
| Mistral 7B | 65 | 61 | 30 |

TABLE VII: Counts of trigger hallucinations across all experiments.

Table VII shows hallucination counts across experiments. LLAMA and Mistral show higher rates than GPT-4o, but overall counts remain low relative to test scenarios examined, confirming the validation protocol’s effectiveness.

H. Cost Analysis

We explicitly compare WIT against all baseline categories in Table III, covering inference latency, API cost, and compute requirements across conventional supervised and LLM-based supervised settings. Conventional supervised baselines such as XGBoost-EC [9] and EFR-ERC_{multi} [2] have a cost advantage at inference as they require no LLM calls and operate on lightweight architectures, though they require complete utterance-level annotation at training. Both WIT and the LLM-based supervised baseline Stacked Instruction Tuning [8] require two model calls at inference, one for question generation and one for prediction, resulting in comparable latency (mean 2.1s for WIT and 1.9s for Stacked Instruction Tuning per conversation, using GPT-4o). Regarding cost, GPT-4o-based WIT incurs marginally higher API cost than the supervised baseline due to longer input length from the question series. Open-source WIT models (LLAMA 3 8B, Mistral 7B) eliminate API costs while remaining competitive in performance with GPT-4o-based WIT, though self-hosted deployment requires GPU compute resources.

WIT reduces annotation time by $\sim 60\%$ during training through shift-boundary labelling versus complete utterance annotation. Supervised methods involve two stages: first predicting utterance-level emotion labels, then predicting emotion shifts based on those predictions. Utterance-level emotion annotation is required at training, and during inference, prediction errors from the first stage can propagate to shift prediction. In contrast, WIT directly predicts shifts at inference with zero annotation overhead. WIT achieves F1 scores close to supervised SOTA across datasets: 0.749 vs 0.760 on MELD-FR (gap: 0.011) and 0.783 vs 0.790 on EFR-MASAC (gap: 0.007), demonstrating competitive performance with substantially reduced supervision. Quality control on generated instructions (96% validation accuracy, Section III-A) ensures reliable training data. While supervised methods remain optimal when complete annotations are available, WIT provides a practical and effective alternative when utterance-level labels are scarce or costly to obtain.

I. Limitations

Several limitations warrant discussion. First, generated wayfinding instructions may contain errors; while spot-validation shows 96% accuracy, systematic verification across all training data was not conducted. Second, our evaluation

focuses on Ekman's discrete emotions in English and code-mixed dialogues; extensions to other emotion taxonomies, languages, or cultural contexts require further validation. Third, emotion shift triggers in real conversations involve complex factors beyond textual utterances, including tone, timing, shared history, and cultural context, which our text-only approach cannot fully capture. These limitations suggest directions for future multimodal and cross-cultural research.

V. CONCLUSION

This study demonstrates the effectiveness of LLMs in identifying emotion shift triggers in multi-party conversations provided they are tuned with high-quality, context-sensitive instructions. Technically, we propose a structured wayfinding prompting strategy that systematically decomposes the ESR task into targeted sub-questions for instruction tuning. Leveraging the WIT framework, frontier models such as GPT-4o and open-source models such as LLAMA 3 8B, and Mistral 7B excel when instruction-tuned with dynamic wayfinding questions that emulate human reasoning. Our experiments show performance comparable to SOTA benchmarks with substantially reduced annotation requirements. Building upon this, future work should focus on improving generalisability by expanding experiments to capture nuanced emotional blends, and incorporating multimodal cues such as vocal tones and gestures. These advancements would further enhance the scalability of emotion-aware systems.

REFERENCES

- [1] G. Gamage, D. De Silva, N. Mills, D. Alahakoon, and M. Manic, "Emotion aware: an artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis," *Journal of Big Data*, vol. 11, no. 1, Jul. 2024. [Online]. Available: <http://dx.doi.org/10.1186/s40537-024-00953-2>
- [2] S. Kumar, A. Shrimal, M. S. Akhtar, and T. Chakraborty, "Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer," *Knowledge-Based Systems*, vol. 240, p. 108112, Mar. 2022. [Online]. Available: <http://dx.doi.org/10.1016/j.knosys.2021.108112>
- [3] Y. Khurana, S. Gupta, R. Sathyaraj, and S. P. Raja, "Robinnet: A multimodal speech emotion recognition system with speaker recognition for social interactions," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 1, p. 478–487, Feb. 2024. [Online]. Available: <http://dx.doi.org/10.1109/TCSS.2022.3228649>
- [4] Z. Zhang, S. Wang, W. Yu, Y. Xu, D. Iter, Q. Zeng, Y. Liu, C. Zhu, and M. Jiang, "Auto-instruct: Automatic instruction generation and ranking for black-box language models," *ACM*, 2023. [Online]. Available: <http://dx.doi.org/10.18653/v1/2023.findings-emnlp.659>
- [5] W. Li, Y. Li, V. Pandelea, M. Ge, L. Zhu, and E. Cambria, "Espec: Emotion-cause pair extraction in conversations," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, p. 1754–1765, Jul. 2023. [Online]. Available: <http://dx.doi.org/10.1109/taffc.2022.3216551>
- [6] J. Pan, R. Liang, Z. He, J. Li, Y. Liang, X. Zhou, Y. He, and Y. Li, "St-scgnn: A spatio-temporal self-constructing graph neural network for cross-subject eeg-based emotion recognition and consciousness detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, p. 777–788, Feb. 2024. [Online]. Available: <http://dx.doi.org/10.1109/JBHI.2023.3335854>
- [7] S. Kumar, M. S. Akhtar, E. Cambria, and T. Chakraborty, "Semeval 2024 - task 10: Emotion discovery and reasoning its flip in conversation (ediref)." *ACM*, 2024, p. 1933–1946. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.semeval-1.270>
- [8] V. Nguyen and X. Zhang, "Gavx at semeval-2024 task 10: Emotion flip reasoning via stacked instruction finetuning of llms." *ACM*, 2024, p. 326–336. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.semeval-1.50>
- [9] W. Tian, P. Ji, L. Zhang, and Y. Jian, "Tw-nlp at semeval-2024 task10: Emotion recognition and emotion reversal inference in multi-party dialogues." *ACM*, 2024, p. 311–315. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.semeval-1.48>
- [10] N. Wan, S. Au, E. Ubale, and D. Krogh, "Ucsc nlp at semeval-2024 task 10: Emotion discovery and reasoning its flip in conversation (ediref)." *ACM*, 2024, p. 1492–1497. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.semeval-1.214>
- [11] S. Patel, D. Shukla, and A. Modi, "Iitk at semeval-2024 task 10: Who is the speaker? improving emotion recognition and flip reasoning in conversations via speaker embeddings." *ACM*, 2024, p. 1811–1820. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.semeval-1.256>
- [12] A. Shanbhag, S. Jadhav, S. Rathi, S. Pande, and D. Kadam, "Innovators at semeval-2024 task 10: Revolutionizing emotion recognition and flip analysis in code-mixed texts." *ACM*, 2024, p. 634–641. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.semeval-1.93>
- [13] J. Wang and T. Mine, "Multi-task learning for emotion recognition in conversation with emotion shift," C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li, and J. Li, Eds. Hong Kong, China: ACM, Dec. 2023, pp. 257–266. [Online]. Available: <https://aclanthology.org/2023.paclic-1.26>
- [14] N. Lubis, S. Sakti, G. Neubig, K. Yoshino, T. Toda, and S. Nakamura, "A study of social-affective communication: Automatic prediction of emotion triggers and responses in television talk shows," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 777–783.
- [15] R. Lou, K. Zhang, and W. Yin, "Large language model instruction following: A survey of progresses and challenges," *Computational Linguistics*, vol. 50, no. 3, p. 1053–1095, 2024. [Online]. Available: http://dx.doi.org/10.1162/coli_a_00523
- [16] Arefa, M. A. Ansari, C. Saxena, and T. Ahmad, "JMI at SemEval 2024 task 3: Two-step approach for multimodal ECAC using in-context learning with GPT and instruction-tuned llama models," A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds. Mexico City, Mexico: ACM, Jun. 2024, pp. 1561–1576. [Online]. Available: <https://aclanthology.org/2024.semeval-1.223>
- [17] Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," 2024. [Online]. Available: <https://arxiv.org/abs/2406.11161>
- [18] F. Xu and G. Li, "Multiple information extraction and interaction for emotion recognition in multi-party conversation," in *2023 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, Jul. 2023, p. 977–982. [Online]. Available: <http://dx.doi.org/10.1109/ISCC58397.2023.10218135>
- [19] Y. Li, Y. Gong, C.-H. H. Yang, P. Bell, and C. Lai, "Revise, reason, and recognize: Llm-based emotion recognition via emotion-specific prompts and asr error correction," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP49660.2025.10888591>
- [20] X. Hong, Y. Gong, V. Sethu, and T. Dang, "Aer-llm: Ambiguity-aware emotion recognition leveraging large language models," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP49660.2025.10888198>
- [21] Y. Mao, "Emotion-aware systems for hazardous operations: Integrating mobile sensors with llm," in *2024 6th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*. IEEE, Dec. 2024, p. 913–916. [Online]. Available: <http://dx.doi.org/10.1109/RICAI64321.2024.10911410>
- [22] S. Bo-Hao, S. G. Upadhyay, and L. Chi-Chun, "Toward zero-shot speech emotion recognition using llms in the absence of target data," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP49660.2025.10889305>
- [23] S. Dutta and S. Ganapathy, "Llm supervised pre-training for multimodal emotion recognition in conversations," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP49660.2025.10889998>
- [24] P. Müller, A. Heimerl, S. M. Hossain, L. Siegel, J. Alexandersson, P. Gebhard, E. André, and T. Schneberger, "Recognizing emotion regulation strategies from human behavior with large language models," in *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Sep. 2024, p. 210–218. [Online]. Available: <http://dx.doi.org/10.1109/ACII63134.2024.00029>

- [25] Y. Zhang, H. Liu, Y. Xiao, M. Amoon, D. Zhang, D. Wang, S. Yang, and C. Quek, "Llm-enhanced multi-teacher knowledge distillation for modality-incomplete emotion recognition in daily healthcare," *IEEE Journal of Biomedical and Health Informatics*, p. 1–11, 2024. [Online]. Available: <http://dx.doi.org/10.1109/JBHI.2024.3470338>
- [26] L. Peng, Z. Zhang, T. Pang, J. Han, H. Zhao, H. Chen, and B. W. Schuller, "Customising general large language models for specialised emotion recognition tasks," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2024, p. 11326–11330. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP48485.2024.10447044>
- [27] C. Yang, Y. Zhu, W. Lu, Y. Wang, Q. Chen, C. Gao, B. Yan, and Y. Chen, "Survey on knowledge distillation for large language models: Methods, evaluation, and application," *ACM Transactions on Intelligent Systems and Technology*, Oct. 2024. [Online]. Available: <http://dx.doi.org/10.1145/3699518>
- [28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2022. [Online]. Available: <https://arxiv.org/abs/2201.11903>
- [29] Z. Chu, J. Chen, Q. Chen, W. Yu, T. He, H. Wang, W. Peng, M. Liu, B. Qin, and T. Liu, "Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future." *ACM*, 2024, p. 1173–1203. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.acl-long.65>
- [30] D. Hershovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. Cabello Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust, and A. Søgaard, "Challenges and strategies in cross-cultural nlp." *ACM*, 2022. [Online]. Available: <http://dx.doi.org/10.18653/v1/2022.acl-long.482>
- [31] A. Madaan, K. Hermann, and A. Yazdanbakhsh, "What makes chain-of-thought prompting effective? a counterfactual study." *ACM*, 2023, p. 1448–1535. [Online]. Available: <http://dx.doi.org/10.18653/v1/2023.findings-emnlp.101>
- [32] M. Amirizani, E. Martin, M. Sivachenko, A. Mashhadi, and C. Shah, "Can llms reason like humans? assessing theory of mind reasoning in llms for open-ended questions," ser. CIKM '24. *ACM*, Oct. 2024, p. 34–44. [Online]. Available: <http://dx.doi.org/10.1145/3627673.3679832>
- [33] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction tuning for large language models: A survey," 2023. [Online]. Available: <https://arxiv.org/abs/2308.10792>
- [34] S. Dabhi, J. Martinez, and F. Poursardar, "Evaluating the effectiveness of fine-tuning large language model for domain-specific task," in *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, Aug. 2024, p. 176–177. [Online]. Available: <http://dx.doi.org/10.1109/IRI62200.2024.00045>
- [35] J. Liu, Z. Kong, P. Dong, X. Shen, P. Zhao, H. Tang, G. Yuan, W. Niu, W. Zhang, X. Lin, D. Huang, and Y. Wang, "Rora: Efficient fine-tuning of llm with reliability optimization for rank adaptation," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP49660.2025.10889613>
- [36] S. Hao and L. Duan, "Online learning from strategic human feedback in llm fine-tuning," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP49660.2025.10887891>
- [37] H. Govindharajan and S. Vijayakumar, "A framework for automated selective fine-tuning of domain-specific large language models using graph-based retrieval augmented generation," in *2024 IEEE 15th Annual Ubiquitous Computing, Electronics amp; Mobile Communication Conference (UEMCON)*. IEEE, Oct. 2024, p. 431–439. [Online]. Available: <http://dx.doi.org/10.1109/UEMCON62879.2024.10754778>
- [38] K. Schlegel, N. R. Sommer, and M. Mortillaro, "Large language models are proficient in solving and creating emotional intelligence tests," *Communications Psychology*, vol. 3, no. 1, p. 80, 2025.
- [39] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu, "Emotional intelligence of large language models," *Journal of Pacific Rim Psychology*, vol. 17, p. 18344909231213958, 2023.
- [40] S. Sabour, S. Liu, Z. Zhang, J. Liu, J. Zhou, A. Sunaryo, T. Lee, R. Mihalcea, and M. Huang, "Emobench: Evaluating the emotional intelligence of large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 5986–6004.
- [41] P. E. Griffiths, "Basic emotions, complex emotions, machiavellian emotions," March 2002. [Online]. Available: <https://philsci-archives.pitt.edu/604/>
- [42] F. M. Plaza-del Arco, A. A. Cercas Curry, A. Cercas Curry, and D. Hovy, "Emotion analysis in NLP: Trends, gaps and roadmap for future directions," N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 5696–5710. [Online]. Available: <https://aclanthology.org/2024.lrec-main.506>
- [43] K. K. N. R. Verma, and J. Shukla, "Towards context-aware eeg-based emotion recognition models: Personality and emotional intelligence as context," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP49660.2025.10890528>
- [44] G. M. Caucci and R. J. Kreuz, "Social and paralinguistic cues to sarcasm," *Humor*, vol. 25, no. 1, p. 1–22, Jan. 2012. [Online]. Available: <http://dx.doi.org/10.1515/humor-2012-0001>
- [45] A. Hassan, N. A. Helal, Y. M. Afify, and N. L. Badr, "Exploring emotional cues within conversational contexts for sarcasm detection," in *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. IEEE, Oct. 2024, p. 127–130. [Online]. Available: <http://dx.doi.org/10.1109/NILES63360.2024.10753268>
- [46] D. Ghosal, N. Majumder, R. Mihalcea, and S. Poria, "Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study." *ACM*, 2021, p. 1435–1449. [Online]. Available: <http://dx.doi.org/10.18653/v1/2021.findings-acl.124>
- [47] S. Brown-Schmidt, S. O. Yoon, and R. A. Ryskin, *People as Contexts in Conversation*. Elsevier, 2015, p. 59–99. [Online]. Available: <http://dx.doi.org/10.1016/bs.plm.2014.09.003>
- [48] C. A. Smith and P. C. Ellsworth, "Patterns of cognitive appraisal in emotion," *Journal of Personality and Social Psychology*, vol. 48, no. 4, p. 813–838, 1985. [Online]. Available: <http://dx.doi.org/10.1037/0022-3514.48.4.813>
- [49] T. D. Kemper and R. S. Lazarus, "Emotion and adaptation." *Contemporary Sociology*, vol. 21, no. 4, p. 522, Jul. 1992. [Online]. Available: <http://dx.doi.org/10.2307/2075902>
- [50] G. C. Yeo and D. C. Ong, "Associations between cognitive appraisals and emotions: A meta-analytic review." *Psychological Bulletin*, vol. 150, no. 12, p. 1440–1471, Dec. 2024. [Online]. Available: <http://dx.doi.org/10.1037/bul0000452>
- [51] Q. Gao, B. Cao, X. Guan, T. Gu, X. Bao, J. Wu, B. Liu, and J. Cao, "Emotion recognition in conversations with emotion shift detection based on multi-task learning," *Knowledge-Based Systems*, vol. 248, p. 108861, Jul. 2022. [Online]. Available: <http://dx.doi.org/10.1016/j.knosys.2022.108861>
- [52] G. Tu, F. Xiong, B. Liang, and R. Xu, "A persona-infused cross-task graph network for multimodal emotion recognition with emotion shift detection in conversations," ser. SIGIR 2024. *ACM*, Jul. 2024, p. 2266–2270. [Online]. Available: <http://dx.doi.org/10.1145/3626772.3657944>
- [53] S. Kumar, M. S. Akhtar, E. Cambria, and T. Chakraborty, "Semeval 2024 - task 10: Emotion discovery and reasoning its flip in conversation (ediref)." *ACM*, 2024, p. 1933–1946. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.semeval-1.270>
- [54] Z. H. Shaik, D. Prasanna, E. Jahnvi, R. Thippireddy, V. Madhav, S. Saumya, and S. Biradar, "Feedforward at semeval-2024 task 10: Trigger and sentext-height enriched emotion analysis in multi-party conversations." *ACM*, 2024, p. 745–756. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.semeval-1.107>
- [55] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *ACM Computing Surveys*, vol. 55, no. 9, p. 1–35, Jan. 2023. [Online]. Available: <http://dx.doi.org/10.1145/3560815>
- [56] S. Yao, D. Yu, J. Zhao, I. Shafraan, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," 2023. [Online]. Available: <https://arxiv.org/abs/2305.10601>
- [57] J. Long, "Large language model guided tree-of-thought," 2023. [Online]. Available: <https://arxiv.org/abs/2305.08291>
- [58] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," 2023. [Online]. Available: <https://arxiv.org/abs/2302.00923>
- [59] Y. Zhang, H. Fei, D. Li, and P. Li, "Promptgen: Automatically generate prompts using generative models," in *Findings of the ACM: NAACL 2022*. *ACM*, 2022, p. 30–37. [Online]. Available: <http://dx.doi.org/10.18653/v1/2022.findings-naacl.3>

- [60] K. Shum, S. Diao, and T. Zhang, "Automatic prompt augmentation and selection with chain-of-thought from labeled data." ACM, 2023. [Online]. Available: <http://dx.doi.org/10.18653/v1/2023.findings-emnlp.811>
- [61] A. Akella, A. Manatkar, B. Chavda, and H. Patel, "An automatic prompt generation system for tabular data tasks." ACM, 2024, p. 191–200. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.naacl-industry.16>
- [62] V. S. Dorbala, S. Chowdhury, and D. Manocha, "Can llm's generate human-like wayfinding instructions? towards platform-agnostic embodied instruction synthesis." ACM, 2024, p. 258–271. [Online]. Available: <http://dx.doi.org/10.18653/v1/2024.naacl-short.24>
- [63] M. Luo, H. Zhang, S. Wu, B. Li, H. Han, and H. Fei, "NUS-emo at SemEval-2024 task 3: Instruction-tuning LLM for multimodal emotion-cause analysis in conversations," A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds. Mexico City, Mexico: ACM, Jun. 2024, pp. 1589–1596. [Online]. Available: <https://aclanthology.org/2024.semeval-1.226>
- [64] H. Hu, S. Yu, P. Chen, and E. M. Ponti, "Fine-tuning large language models with sequential instructions," 2024. [Online]. Available: <https://arxiv.org/abs/2403.07794>
- [65] Y. Qin, Y. Yang, P. Guo, G. Li, H. Shao, Y. Shi, Z. Xu, Y. Gu, K. Li, and X. Sun, "Unleashing the power of data tsunami: A comprehensive survey on data assessment and selection for instruction tuning of language models," 2024. [Online]. Available: <https://arxiv.org/abs/2408.02085>
- [66] P. Ekman, "Basic emotions," p. 45–60, Feb. 1999. [Online]. Available: <http://dx.doi.org/10.1002/0470013494.ch3>
- [67] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, p. 100943–100953, 2019. [Online]. Available: <http://dx.doi.org/10.1109/ACCESS.2019.2929050>
- [68] M. Bedi, S. Kumar, M. S. Akhtar, and T. Chakraborty, "Multi-modal sarcasm detection and humor classification in code-mixed conversations," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, p. 1363–1375, Apr. 2023. [Online]. Available: <http://dx.doi.org/10.1109/TAFFC.2021.3083522>