Deep Emo Seek: Unifying Emotional Intelligence Using Chain of Thought Reasoning

Group 24

Shenghao Yang (122090655)
Xuhuan Shen (123090490)
Chang Su (12109048)
Siqi Yao (122090664)
Tianyu He (121090168)

2025/5/8

Abstract

Modeling emotional intelligence in machines is a challenging yet essential aspect of naturalistic human-computer interaction. Traditional approaches fragment unified emotional intelligence into several tasks and model them separately, which might introduce task-specific biases in separate models and reduces efficiency in practice. In this paper, we propose a unified framework that utilizes Chain-of-Thought (CoT) reasoning to simultaneously handle emotion recognition, reasoning, cause extraction, and dialogue generation. We construct emotion dataset with CoT and use it to fine-tune a Deepseek-R1 model with LoRA under this unified framework. The results demonstrate the benefits of modeling emotional intelligence using the unified framework in interpretability and performance, and pave the way for a holistic emotional intelligence.

1 Introduction

Understanding and modeling human emotions has long been a foundational research challenge across psychology, linguistics, and artificial intelligence. Psychological theories such as *Lazarus' Cognitive Appraisal Theory* [1] and *Affect Control Theory* [2] have provided the conceptual groundwork for modeling human emotional processes. These theoretical frameworks have inspired computational approaches aimed at equipping machines with emotional perception and reasoning capabilities.

However, while recent advances in Large Language Models (LLMs) have showcased impressive capabilities in natural language generation and possibilities of serving as agents for modeling the emotional intelligence, most existing approaches [3–6] remain narrowly focused. Tasks such as emotion recognition, cause extraction, and dialogue generation are typically treated in isolation, resulting in fragmented emotional intelligence and limited practical application in real-world human-computer interaction. Nevertheless, LLMs often operate as passive analysts, lacking the ability to reason through emotional contexts in a manner akin to human thought. Furthermore, existing models rarely account for the dynamic and transitional nature of emotions, which is fundamental to modeling real-world emotional trajectories in dialogue.

To address these limitations, we propose a unified framework that leverages recent advances of Chain-of-Thought (CoT) reasoning (esp. Deepseek-R1) to model emotional intelligence in a structured, interpretable, and holistic manner. CoT allows LLMs to explicitly reason through emotional cues, causes, and shifts over the course of a conversation, bringing together recognition, reasoning, and generation in a coherent process.

Related Works In recent years, LLMs have emerged as powerful tools for emotion-related tasks, prompting a surge in research exploring their emotional competencies. For example, Li et al. [3] investigate direct emotion prompting techniques, while Lei et al. [5] propose instruction-tuned models for Emotion Recognition in Conversation (ERC). Other notable works include Emotion-CoT [4], which introduces step-by-step emotion reasoning, and ESCOT

[6], which enhances emotional support dialogue generation via CoT-style prompts. Collectively, these works highlight the growing potential of LLMs to reason about, recognize, and express emotions in dialogue.

Relevant Datasets A key bottleneck in advancing emotionally intelligent systems is the lack of comprehensive and task-integrated datasets. Most existing datasets are designed for specific subtasks, resulting in a fragmented landscape:

- **Emotion Recognition in Conversation (ERC):** Datasets such as *IEMOCAP* [7], *MELD* [8], and *EmoryNLP* [9] provide annotated dialogues for recognizing speaker emotions. These corpora focus primarily on emotion classification but do not support cause inference or generative emotional reasoning.
- **Emotion Cause Pair Extraction (ECPE):** Datasets like *RECCON* [10] and *ConvECPE* [11] offer annotations for extracting emotion-cause pairs from text. However, they often lack dialogue structure and contextual emotion evolution.
- **Emotional Reasoning and Multi-Task Dialogues:** Emerging datasets such as *CICEROv2* [12], *EMER* [13], and *EDEN* [14] attempt to model deeper emotion reasoning, including moral appraisal and empathy. *DailyDialogue* [15] spans both ERC and Emotional Dialogue Generation (EDG), but remains limited in terms of causal annotations and reasoning depth.

Our key contributions in this paper are:

- A CoT-based reasoning framework that unifies multiple emotional tasks, including recognition, cause inference, shift detection, and generation.
- A novel dataset extending MELD with structured CoT annotations that provide interpretable reasoning paths for emotional inference.
- A fine-tuned Deepseek-R1 model using Low-Rank Adaptation (LoRA) that demonstrates strong performance on CoT-based emotional tasks with high efficiency.

2 Unified Framework for Emotional Intelligence Modeling

To model emotional intelligence comprehensively, we propose a unified Chain-of-Thought (CoT) based reasoning framework that consolidates four key stages for emotional intelligence: emotion perception, emotion cognition, emotion expression, and emotion reflection. These tasks reflect the multilayered nature of emotional understanding in dialogue, as illustrated in *Fig.1.* In the emotion perception stage, the emotional intelligence should be capable of perceiving other's emotions. Upon understanding other's emotions, in the emotion perception stage, a cognitive process will lead to the discovery of self-emotion. Then, emotion would be expressed in the emotion expression stage. Finally, the emotional intelligence would reflect upon its thinking process to understand the flow of emotion. All four stages are not isolated but connected with one CoT, which chains all the tasks together in a way that resembles the sequential and logical cognitive process in human emotional intelligence while maintaining the interpretability.

CoT prompting [16] allows models to decompose complex tasks into a series of interpretable reasoning steps. Instead of producing a final answer directly, the model walks through intermediate inferences such as identifying context, emotional cues, and potential causes before concluding the emotional state. This approach is particularly effective in emotion-related tasks where ambiguity and contextual nuance are prevalent.

By integrating CoT with Deepseek-R1, our model achieves both high accuracy and interpretability. The CoT structure also supports generative outputs: in addition to predicting emotion labels, the model generates final utterances for emotional dialogues and produces explanatory rationales. This generative aspect is crucial for emotion reasoning and reflective dialogue tasks.

The framework is evaluated in several metrics and generation tasks. The model needs to performs reasoning to process emotional context, predict emotional transitions, and generate both the final utterance and a rationale for the speaker's emotional state. All metrics used includes:

- **Previous Emotion**: Emotion label of the speaker's most recent prior utterance.
- **Previous Counterpart Emotion**: Emotion label of the most recent utterance by another speaker in the dialogue.
- **Emotion Shift**: Whether and how the speaker's emotion changes over time.
- **Final Emotion**: The emotion expressed in the current utterance.
- **Emotion Cause**: The inferred cause or trigger for the emotional state or shift.

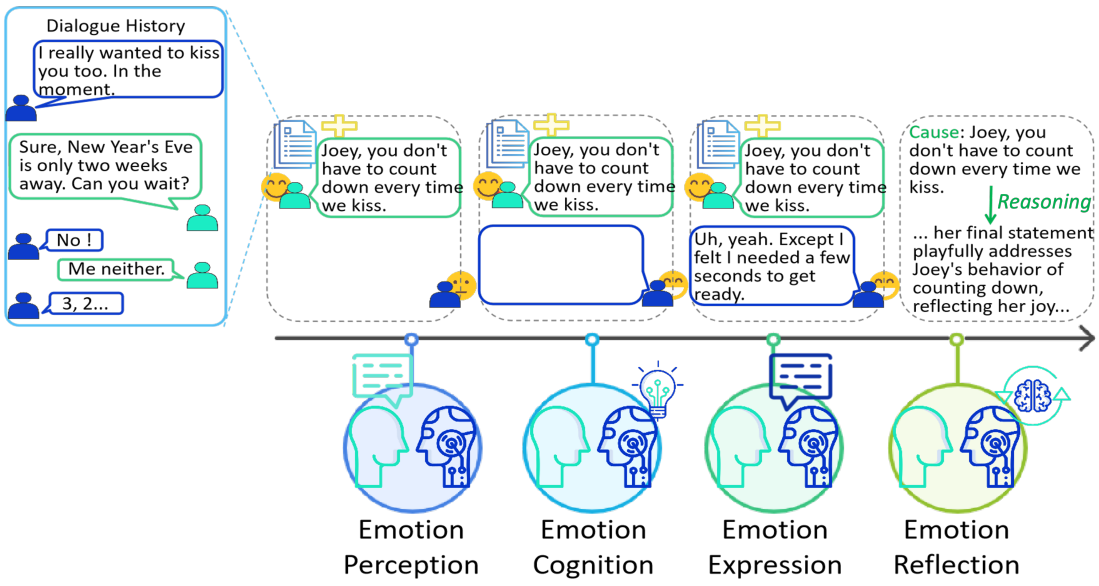

Figure 1: Unified Chain-of-Thought Framework for Emotional Intelligence Tasks

Please note that the task is, in essence, a generative task. Explicit labels extracted from the response of the model are evaluated as listed in above. The evaluation of reasoning task can also be evaluated by feeding the generated rationales into other advanced reasoning models to check whether the generated logical chains are plausible.

3 Dataset Construction with CoT Ground Truth

The framework is evaluated on the MELD dataset [8], which provides lines of the TV show *Friends* and corresponding emotion label for each line (samples of the dataset can be found in *Fig.2*).

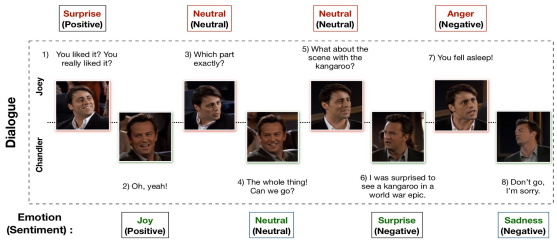

Figure 2: samples with emotion labels from MELD dataset

To enable emotion reasoning by fine-tuning reasoning models with Chain-of-Thought (CoT) supervision, we construct a CoT-augmented version of the aforementioned MELD dataset so each data instance is annotated not only with emotion labels but also with intermediate reasoning steps that explain how the model should arrive at those labels based on dialogue context.

To attain CoT ground-truth, our initial strategy was to directly generate CoT reasoning using Deepseek-R1 from raw dialogue, without providing the ground-truth emotion label. However, this approach failed: the model frequently produced incorrect emotion predictions, leading to faulty reasoning chains. Since the goal is to produce valid reasoning steps, starting from an incorrect answer inherently hurt the quality and reliability of the CoT annotations.

In a second attempt, we provided both the dialogue and the ground-truth label to R1 and asked it to generate a CoT reasoning path. While this method yielded better results in terms of label accuracy, the model often relied too heavily on the given label. Many outputs began with phrases such as "Given that the emotion is X..." or "From the answer, we can infer...". However, the CoT containing direct reference to the answer is misleading to the model training and should not be used as the guidance of training, as there would not be answers available in testing scenarios. Thus, labels can not be a direct guidance to generate the CoT groundtruths [17].

To address the limitations of the above naive methods, we devised a more principled approach. The core philosophy is to leverage CoT steps, simplified CoT, instead of labels, to serve as the guidance to generate the final CoT groundtruth. The designed CoT steps

are structural and informative enough to be logical to serve as the guidance to generate the CoT, but short enough to avoid being directly and misleadingly referred in generated CoT. To achieve these goals, our pipeline involves four key stages (*Fig.3*):

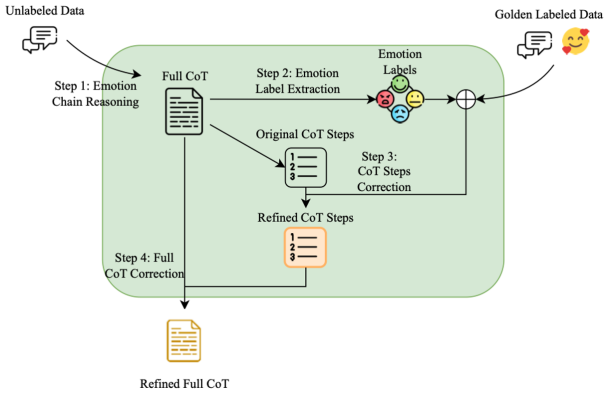

Figure 3: Pipeline for CoT dataset construction

- **CoT Inference**: We take the raw dialogue without any emotion labels and feed it into DeepSeek-R1. The model generates a complete chain-of-thought for each utterance, detailing inferred emotional states, causal factors, and contextual interpretations, but this initial reasoning often contains mistakes or omissions in reasoning.
- **Labels Extraction**: The CoT output generated in CoT Inference step is passed to DeepSeek-V3, which parses out the emotion labels. As the generated CoT is often incorrect, these labels are also often incorrect, but they are useful in the later step to serve as the anchors to find the mistakes.
- **Attaining CoT Steps**: We combine the groundtruth labels provided in MELD dataset, the extracted labels from incorrect generated CoT, the original dialogue, and the full generated CoT from Step 1, and re-prompt DeepSeek-R1. This time, R1 is instructed to produce only a concise sequence of CoT reasoning steps, essentially a skeletal outline of the thought process. The result is a refined list of intermediate reasoning steps.
- **Rewrite CoT with Guided Steps**: Finally, as we have got the correct CoT steps in last step and can now use them as the guidance to generate the correct full CoT, we merge the refined CoT steps with the generated mistaken full CoT from Step 1 and run them through DeepSeek-V3 once more. The model uses the distilled step outline to produce a polished Chain-of-Thought and corresponding labels.

To ensure that the generated CoT in Step 4 is valid, we evaluated the labels from step 4, and found that in 90% samples, all metrics are fully correct. We future re-prompt DeepSeek-V3 to rewrite the full CoT with additional information about which metrics are incorrect. After 2 rounds of re-prompting, in 95 samples, all metrics are fully correct. This indicates that the corresponding CoT are also well grounded. Additionally, we also manually check the final CoT, which validates that the CoT is very logical and inherent with the context.

4 Experiments

4.1 Model Training and Fine-Tuning

We adopt **Deepseek-R1-distill-Qwen7B** as our base model. This variant is a distilled and lighter-weight version of Deepseek-R1, consisting of approximately 7 billion parameters. It offers a favorable trade-off between computational efficiency and reasoning power.

To enable robust emotional inference, we fine-tune the model on both CoT reasoning parts and final outputs. The training is conducted on the CoT-augmented MELD dataset described in the previous section.

To achieve efficient adaptation, we apply **Low-Rank Adaptation (LoRA)** [18]. LoRA introduces trainable rank decomposition matrices into the attention weights of the transformer architecture, enabling low-resource fine-tuning. Specifically, the adaptation to the pretrained weights is performed as follows:

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

where W_0 is the frozen original weight, and $\Delta W = BA$ is the learned low-rank update, with $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ for a chosen rank r.

The training objective is to maximize the log-likelihood over our CoT-augmented dataset:

$$\max_{\Theta} \sum_{(x,y) \in \mathcal{Z}} \sum_{t=1}^{|y|} \log \left(p_{\Phi_0 + \Delta\Phi(\Theta)}(y_t \mid x, y_{<t}) \right)$$

where Φ_0 represents the frozen base model parameters, and $\Delta\Phi(\Theta)$ represents the LoRA trainable parameters.

Table 1: Emotion Analysis Performance Comparison

Model	Prev. Emo. F1	Prev. Cptr. Emo. F1	Emo. Shift Acc	Final Emo. F1	Emo. Cause F1	Emo. Reasoning
GPT-4o	0.003	0.004	0.002	0.003	0.005	0.007
Deepseek-R1 (671B)	0.486	0.495	0.535	0.237	**0.711**	1.242
Finetuned Qwen2-7B	0.527	0.419	0.561	0.232	0.571	1.162
Deep Emo Seek (Finetuned DeepSeek-R1-distill-Qwen-7B)	**0.647**	**0.594**	**0.606**	**0.362**	0.586	**1.323**

Implementation Details All models were trained on a single workstation equipped with four NVIDIA RTX 4090 GPUs. We apply LoRA to every transformer module, instantiating each low-rank decomposition with a rank of 8 and a scaling factor (α) of 16, and regularize with a dropout rate of 0.1. We optimize using a constant learning rate of 5×10^{-5} throughout training.

Addtionally, we aconduct experiments using GPT-4o and DeepSeek-R1 (671B) without finetuning for baselines, and we finetune Qwen2-7B with the same method to compare reasoning models (R1) and non-reasoning models.

4.2 Results

Model evaluation is based on the emotional task metrics introduced above, including accuracy, F1 scores, and qualitative coherence of reasoning for the following subtasks: previous emotion, counterpart emotion, emotion shift, final emotion, and emotion cause (Tab. 1).

Key Findings Our experiments reveal several important insights:

- **State-of-the-art performance:** Deep Emo Seek achieves the highest F1 scores on previous-emotion (0.647), counterpart-emotion (0.594) and emotion-shift accuracy (0.606), outperforming both the vanilla reasoning model (Deepseek-R1) and the non-reasoning finetuned model (Qwen2-7B).

- **Enhanced final reasoning:** On the challenging "final emotion" task, Deep Emo Seek improves over Deepseek-R1 by 52%, demonstrating that our distillation and guided-step approach sharpens outcome prediction.

- **Poor zero-shot results:** GPT-4o in zero-shot mode fails to capture any meaningful emotional signals, with F1 and accuracy metrics close to 0, underscoring the necessity of specialized tuning for emotional reasoning.

5 Conclusion

In this work, we present a unified framework for modeling emotional intelligence by leveraging Chain-of-Thought (CoT) reasoning. Our approach integrates multiple emotion-related tasks—emotion recognition, cause extraction, shift detection, reasoning, and dialogue generation—into a single coherent structure. We construct a CoT-augmented version of the MELD dataset that enables step-by-step emotional inference and corrects naive annotation strategies to align with human-like reasoning processes.

To demonstrate the effectiveness this framework, we fine-tune Deepseek-R1-distill-Qwen7B using Low-Rank Adaptation (LoRA) for both CoT reasoning and task-specific outputs. Our empirical results show that this fine-tuned model achieves superior performance across most metrics compared to even larger models without fine-tuning (Deepseek-R1 (671B)) in our framework. Notably, our model performs well despite its significantly smaller parameter count, demonstrating the value of CoT-guided supervision and efficient fine-tuning in this unfied framework.

Beyond performance, our framework contributes to the interpretability and robustness of emotional reasoning in dialogue systems. The modular design and CoT annotations enable future researchers to plug in additional tasks or extend the reasoning chain to model more complex affective phenomena.

References

[1] R. S. Lazarus, "Progress on a cognitive-motivational-relational theory of emotion," *American psychologist*, vol. 46, no. 8, p. 819, 1991.

[2] D. R. Heise, "Affect control theory: Concepts and model," in *Analyzing Social Interaction: Advances in Social Science Research Using R*, pp. 1–33, Springer, 2016.

[3] C. Li, J. Wang, Y. Zhang, Y. Chen, Y. Shen, L. Zhang, Y. Liu, Z. Zhang, and X. Qiu, "Large language models understand and can be en-

hanced by emotional stimuli," *arXiv preprint arXiv:2307.11760*, 2023.

[4] Z. Li, G. Chen, R. Shao, T. Zhang, and R. Xia, "Enhancing emotional generation capability of large language models via emotional chain-of-thought," *arXiv preprint arXiv:2401.06836*, 2024.

[5] S. Lei, G. Dong, X. Wang, X. Chen, W. Zhang, and M. Zhang, "Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework," *arXiv preprint arXiv:2309.11911*, 2023.

[6] T. Zhang, X. Zhang, J. Zhao, T. Liu, and R. Xia, "Escot: Towards interpretable emotional support dialogue systems," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 13395–13412, 2024.

[7] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[8] S. Poria, D. Hazarika, N. Majumder, A. Zadeh, and E. Cambria, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

[9] S. M. Zahiri and J. D. Choi, "Emotion detection on tv show transcripts with sequence-based convolutional neural networks," in *Workshops at the thirty-second aaai conference on artificial intelligence*, 2018.

[10] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya, *et al.*, "Recognizing emotion cause in conversations," *Cognitive Computation*, vol. 13, pp. 1317–1332, 2021.

[11] W. Li, Y. Li, V. Pandelea, M. Ge, L. Zhu, and E. Cambria, "Ecpec: Emotion-cause pair extraction in conversations," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1754–1765, 2022.

[12] S. Shen, D. Ghosal, N. Majumder, H. Lim, R. Mihalcea, and S. Poria, "Multiview contextual commonsense inference: A new dataset and task," *arXiv preprint arXiv:2210.02890*, 2022.

[13] Z. Lian, L. Sun, M. Xu, H. Sun, K. Xu, Z. Wen, S. Chen, B. Liu, and J. Tao, "Explainable multimodal emotion reasoning," *arXiv preprint arXiv:2306.15401*, 2023.

[14] J. Li, Z. Lin, L. Wang, Q. Si, Y. Cao, M. Yu, P. Fu, W. Wang, and J. Zhou, "Think out loud: Emotion deducing explanation in dialogues," *arXiv preprint arXiv:2406.04758*, 2024.

[15] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, 2017.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems (NeurIPS 2022)*, vol. 36, 2022. Conference on Neural Information Processing Systems (NeurIPS 2022).

[17] A. Y. Tsai, A. Kraft, L. Jin, C. Cai, A. Hosseini, T. Xu, Z. Zhang, L. Hong, E. H. Chi, and X. Yi, "Leveraging LLM reasoning enhances personalized recommender systems." arXiv:2408.00802 [cs.IR], Aug 2024. To appear in ACL 2024.

[18] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.