Enhancing Emotional Generation Capability of Large Language Models via Emotional Chain-of-Thought

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown remarkable performance in various emotion recognition tasks, thereby piquing the research community's curiosity for exploring their potential in emotional intelligence. However, several issues in the field of emotional generation tasks remain unresolved, including human preference alignment and emotional generation assessment. In this paper, we propose the Emotional Chain-of-Thought (ECoT), a plug-and-play prompting method that enhances the performance of LLMs on various emotional generation tasks by aligning with human emotional intelligence guidelines. To assess the reliability of ECoT, we propose an automated model-based evaluation method called Emotional Generation Score (EGS). EGS incorporates Goleman's Emotional Intelligence Theory as a consensus of human experts, providing a new perspective on the evaluation of emotional generation tasks. Extensive experimental results demonstrate the effectiveness of ECoT and EGS. Further, we discuss the promise of LLMs in the field of emotional intelligence and present key insights into the LLMs with the ECoT in emotional generation tasks.

1 Introduction

> *Emotional intelligence is a combination of competencies including Self-Awareness, Self-Regulation, Motivation, Empathy and Social Skills.*
>
> - Goleman (1995)

As Large Language Models (LLMs) have shown remarkable capabilities in tasks such as emotion recognition in conversations (Zhang et al., 2023; Lei et al., 2023), aspect-based sentiment analysis (Scaria et al., 2023; Simmering and Huoviala, 2023), emotion classification (Li et al., 2023; Peng et al., 2023), etc., some studies (Wang et al., 2023; Huang et al., 2023; Tanmay et al., 2023) attached their attention to the emotional intelligence of LLMs. Incorporating human emotional intelligence metrics (e.g., MSCEIT (Mayer et al., 2003) and MBTI (Boyle, 1995)) into question-answering tasks enables us to gauge the degree of human-like emotional intelligence exhibited by LLMs.

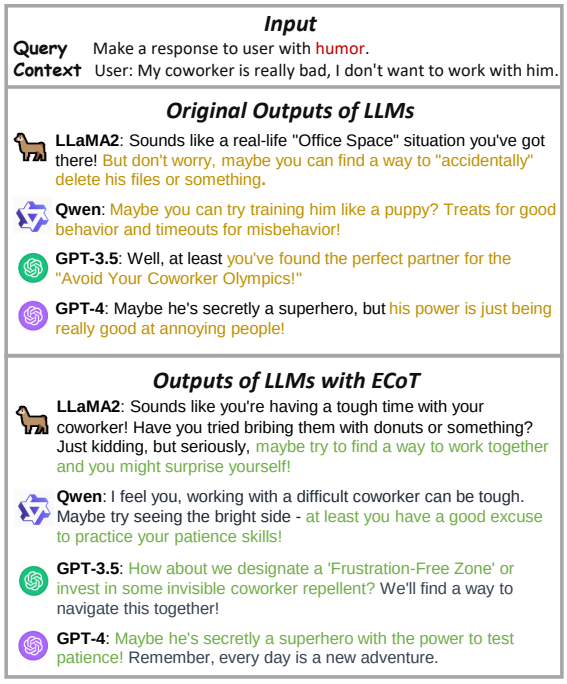


Figure 1: Examples of LLMs generating responses that are harmful to humans. Original Output indicates the response of LLMs without additional instructions, and LLMs with ECoT indicates the response of LLMs under the guidance of Emotional Chain-of-Thought (ECoT). With the query *Make a response to user with humor*, LLMs tend to generate harmful responses, failing to consider the potential negative emotional impact of responses on humans. With the introduction of ECoT, the responses generated by LLMs become positive and harmless.

Nevertheless, it is noteworthy that LLMs continue to entail inherent risks and limitations in the context of emotional generation tasks. As shown in Figure 1, we prompt the LLMs to generate a

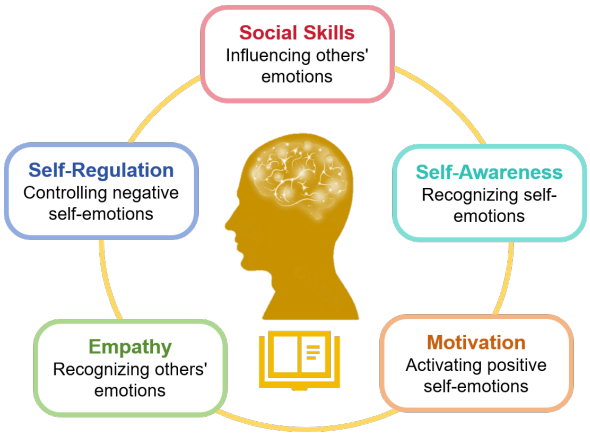


Figure 2: We interpret Goleman's theory through the lens of sentiment analysis and translate them into recognizing self-emotions, controlling negative self-emotions, activating positive self-emotions, recognizing others' emotions, and influencing others' emotions, respectively.

response to *user* using humor as the emotional condition. In the absence of additional instructions, LLMs tend to generate harmful responses that contain disgust, contempt, and mockery of the *coworker*. It reveals that LLMs are deficient in human preference alignment, and ignore the potential emotional impact of responses on humans, which is a less expored area of study. Moreover, it is crucial to acknowledge that emotional generation represents a subjective undertaking. And diverse individuals may possess distinct interpretations and perceptions of emotions. In essence, there is no objective metric to evaluate the performance of LLMs on emotional generation tasks.

With the above problems in mind, we are dedicated to enhancing and evaluating the emotional generation capabilities of LLMs. Inspired by Goleman's theory (Goleman, 1995), we introduce the human emotional intelligence criterion to allow LLMs to align human preferences in emotional generation. Moreover, we propose to assess the quality of emotional generation in terms of multiple dimensions of human emotional intelligence.

Goleman et al. (Goleman, 1995) developed a framework to explain emotional intelligence in terms of five elements: Self-Awareness, Self-Regulation, Motivation, Empathy and Social Skills. More details about the theory will be described in the following sections. In light of the aforementioned theoretical underpinnings, we posit that the emotional generation task could be construed as a multifaceted application of emotional intelligence. It encompasses the ability to discern the emotions of others, adeptly manage one's own emotions, and skillfully navigate interpersonal relationships. Utilizing Figure 1 as an illustrative exemplar, LLMs are required to ascertain the emotional state of the *user*, discern its underlying causes, deliberate upon the potential emotional ramifications of their responses on the *user* and others, then generate responses with appropriate emotion. The absence of these pertinent considerations engenders the propensity of LLMs to produce harmful responses. That is, the response is not aligned to human preferences.

To address the issue of human preference alignment, we propose the Emotional Chain-of-Thought (ECoT) method, which incorporates Goleman's Emotional Intelligence Theory (Goleman, 1995) into the chain-of-thought. The ECoT method operates progressively, leading LLMs through a process of emotion identification, emotional reasoning, and the provision of emotional response. Given the susceptibility of LLMs to input prompts, we furnish ECoT with expert-authored guidelines. It guides the model's cognitive processes towards an emulation of human-like contemplation regarding the emotional content embedded within the context. As shown in Figure 1, LLMs are guided by ECoT to generate responses that become harmless to humans.

To address the issue of emotional generation assessment, we propose an efficient, answer-free evaluation metric, named Emotional Generation Score (EGS). EGS utilizes Goleman's theory as a criterion and introduces GPT-3.5 [1] to evaluate LLMs' responses in multiple dimensions automatically. With the comparisons on four datasets, we confirm that EGS is consistent to human expert ratings. This observation demonstrates the dependability and trustworthiness of the EGS as an evaluation tool for emotional generation.

The contributions of this paper can be summarized as follows:

- We propose the Emotional Chain-of-Thought (ECoT), a plug-and-play prompting method that enhances the performance of LLMs on various emotional generation tasks by aligning it with human preference in emotional intelligence.

[1] https://openai.com/chatgpt

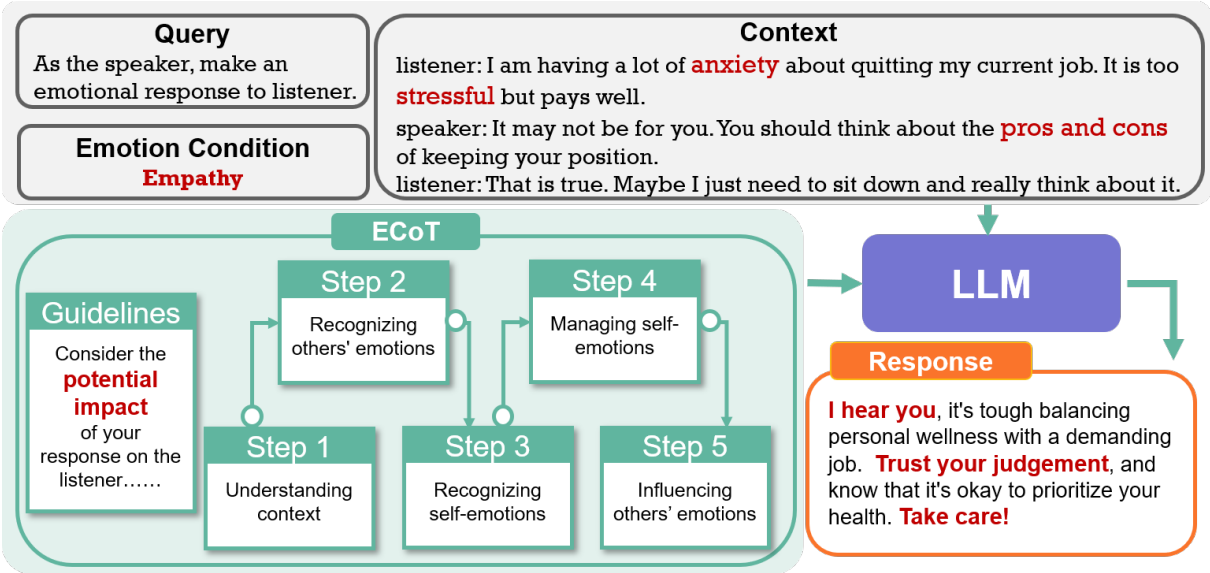


Figure 3: Overview of our proposed ECoT. Given context, emotion condition, task query, and guidelines, LLMs progressively complete understanding context, recognizing others' emotions, recognizing self-emotions, managing self-emotions, influencing others' emotions, then make an emotional response.

- We propose an automatic evaluation method, dubbed EGS, to evaluate the quality of LLMs in emotional conditioned generation. We demonstrate the effectiveness of EGS by comparing it with human expert evaluation.

- In connection with the proposed ECoT, we discuss the prospects for the application of emotion generation. And we present key insights into the LLMs with ECoT in emotional generation tasks.

2 Preliminaries

Here we briefly introduce Goleman's theory (Goleman, 1995) of emotional intelligence, and refer readers to the original paper for more details. Goleman explained emotional intelligence in terms of five elements: Self-Awareness, Self-Regulation, Motivation, Empathy, and Social Skills. Each of these elements is outlined below:

Self-Awareness: checking how your emotions affect your performance; using your values to guide decision-making; learning from your experiences; and being self-confident and certain about your capabilities.

Self-Regulation: controlling your stress by being more positive and action-centered; retaining composure and the ability to think clearly under pressure; handling impulses well; and nurturing trustworthiness and self-restraint.

Motivation: enjoying challenge and stimulation; seeking out achievement; commitment; optimism; and being guided by personal preferences in choosing goals.

Empathy: the ability to see other people's points of view; behaving openly and honestly; and avoiding the tendency to stereotype others.

Social Skills: the use of influencing skills such as persuasion; good communication with others; listening skills; dispute resolution; the ability to inspire and lead others; and the ability to deal with others' emotions.

As shown in Figure 2, we interpreted Goleman's theory through the lens of sentiment analysis and translated the above five dimensions into recognizing self-emotions, controlling negative self-emotions, activating positive self-emotions, recognizing others' emotions, and influencing others' emotions, respectively. Further, we combined the dimensions of controlling negative self-emotions and activating positive self-emotions to managing self-emotions. This serves as the theoretical underpinning for our proposed Emotional Chain-of-Thought (ECoT) methodology.

3 Methodology

3.1 Problem Formulation

Given the context C, denote c^v and c^t as visual context and textual context, respectively. For the context c^t or c^v, emotional generation aims to generate a response R based on emotion condition E

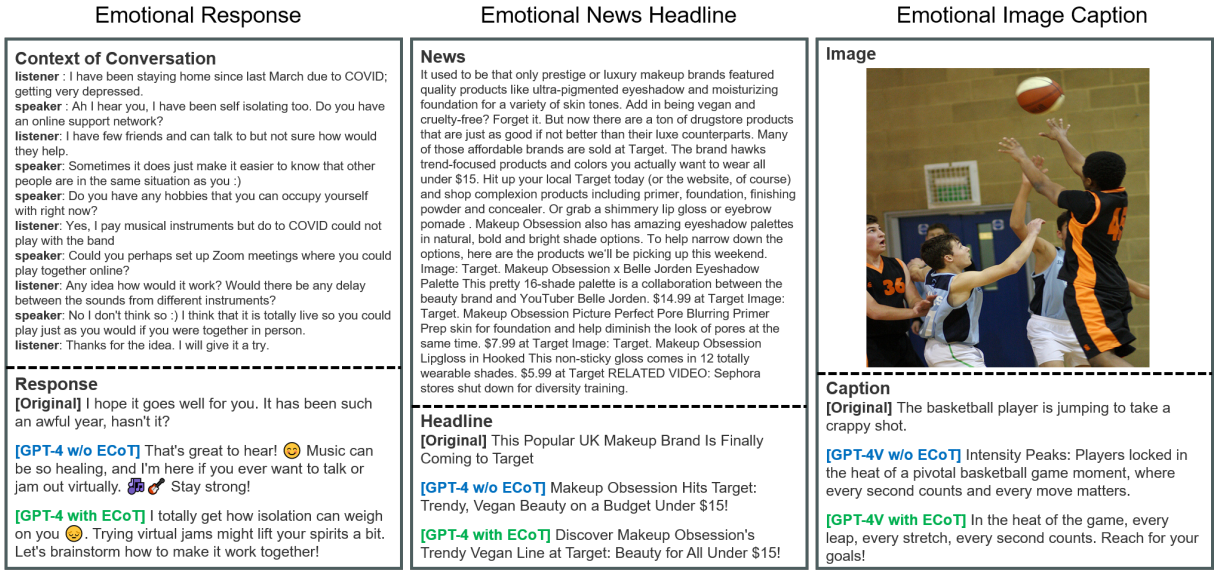


Figure 4: Random samples of emotional generation tasks. The samples of emotional responses, emotional news headline, and emotional image captions were taken from ESConv (Liu et al., 2021a), PENS (Ao et al., 2021a), and SentiCap (Mathews et al., 2016), respectively. *Original* represents the response in the original dataset.

and instruction I.

Emotional generation takes a wide range of forms due to the diversity of context C. Here, we provide a brief overview of some emotional generation tasks.

Emotional Response: Given a two-party conversation as textual context, making a response to the listener from the perspective of the current speaker with a specific emotion (e.g., empathy, happiness, humor, etc.)

Emotional News Headline: Given news as textual context, generating a headline that captures the readers' interest in reading about the news.

Emotional Image Caption: Given an image as visual context, generating a caption that captures the readers' interest in reading about the image.

3.2 Emotional Chain-of-Thought

To explore the emotional generation capabilities of existing LLMs, we conduct a comparative experiment on the IEMOCAP (Busso et al., 2008) dataset. The experimental results demonstrate that various LLMs have limited ability on human preference alignment for emotional generation tasks (see Appendix A for more details).

To address this issue, let's consider how humans make emotional responses in a conversational scenario: initial comprehension of the conversational context, subsequent discernment and comprehension of the emotional state of the interlocutor, and, finally, the formulation of an emotion-specific response. This gives us a key insight: the emotional response is a complex task that is a multifaceted application of emotional intelligence, such as recognizing the emotions of others, controlling one's own emotions, and influencing the emotions of others. It corresponds to Goleman's theory for multiple dimensions of human emotional intelligence. Hence, it is imperative that the LLMs incorporates supplementary emotional thinking processes, thereby enhancing their capacity to emulate the human comprehension and proficient application of emotional intelligence when making emotional responses. In order to implement the emotional thinking process, we suggest incorporating Goleman's theory for multiple dimensions of emotional intelligence into the emotional generation tasks.

Inspired by the remarkable role of Chain-of-Thought (Wei et al., 2022; Wang et al., 2022) in the execution of complex reasoning tasks by LLMs, we introduce Emotional Chain-of-Thought (ECoT) to enable emotional thinking processes in LLMs. Based on Goleman's Theory (Goleman, 1995), we introduce execution procedures that guide the model to perform emotion recognition, emotion reasoning, and emotional generation step by step. As shown in Figure 3, consider the example of LLMs generating emotional responses in a conversation, for a given conversational context C, LLMs perform the following thinking steps:

Step 1: **[Understanding context]** Describe the

Table 1: Experimental results of the baseline models with ECoT on the textual benchmark. The total score for the ratings is 40, and we calculate the average score on each dataset separately. *Original* represents the score of the response in the original dataset.

Model	DailyDialog	IEMOCAP	Empathetic	ESConv	PENS
Original	28.99	25.13	28.26	30.32	23.95
LLama2-13B-Chat	31.32	27.13	22.64	24.47	25.05
+ECoT	**35.20** (↑ 3.88)	**32.42** (↑ 5.29)	**31.29** (↑ 8.65)	**36.52** (↑ 12.05)	**32.90** (↑ 7.85)
Qwen-14B-Chat	30.29	28.87	26.94	27.31	24.02
+ECoT	**34.47** (↑ 4.18)	**32.58** (↑ 3.71)	**32.25** (↑ 5.31)	**35.85** (↑ 8.54)	**32.11** (↑ 8.09)
ChatGLM3-6B	32.01	26.91	28.26	29.44	25.33
+ECoT	**33.33** (↑ 1.32)	**31.52** (↑ 4.61)	**32.19** (↑ 3.93)	**35.56** (↑ 6.12)	**31.66** (↑ 6.33)
GPT-4	34.03	31.62	32.67	33.48	28.53
+ECoT	**36.42** (↑ 2.39)	**36.02** (↑ 4.40)	**36.21** (↑ 3.54)	**36.70** (↑ 3.22)	**33.48** (↑ 4.95)

context of the conversation.

Step 2: **[Recognizing Others' Emotions]** Identify the listener's emotions and explain why.

Step 3: **[Recognizing Self-Emotions]** Identify the speaker's emotions and explain why.

Step 4: **[Managing Self-Emotions]** Consider how to respond in empathy.

Step 5: **[Influencing Others' Emotions]** Consider the impact of response on the listener.

In this way, we explicitly break down the complex emotional response process into a sequential sequence of steps. It not only reveals the "thinking process" of LLMs, but also helps the model decouple the complex task into a collection of simple sub-tasks to improve the performance on emotional generation tasks.

Some works (Zhu et al., 2023; Chang et al., 2023) suggest that LLMs may be sensitive to specific prompts thus affecting their performance. One effective method is to employ in-context learning with LLMs. With the given examples, LLMs learn the form and nature of the task, thus improving the robustness and performance of LLMs. Nevertheless, we found that facilitating the model's acquisition of the fundamental principles of emotional generation through the provision of examples presents a formidable challenge. This challenge arises from two primary factors: (1) unlike objective tasks such as mathematical problem-solving, there are multiple suitable responses for emotional generation tasks. In other words, for a given problem, there are multiple solutions in the answer space; (2) the outcomes of emotional generation tasks are contingent upon the specific context, i.e., a change in context leads to a change in the answer space.

Therefore, instead of offering explicit examples, we suggest incorporating guidelines in ECoT, which outline the expert consensus LLMs must follow. These guidelines are written by human experts and align human preferences in emotional intelligence (see Appendix B for ECoT templates). They function as directives for LLMs, steering them towards to constrained answer spaces as opposed to a specific answer.

To summarize, for a given context C, emotion condition E, task query Q, guidelines G, and thinking steps T, LLMs are expected to output an emotional response R with thinking process P, which can be formulated as the following equation:

$$P, R = LLM_\theta(Q, C, E, G, T). \quad (1)$$

where LLM_θ refers to the LLM with parameters θ.

3.3 Evaluation Method

Emotional generation is a subjective task with human cognitive bias, i.e., *a thousand eyes for a thousand Hamlets*. Thus, methods for evaluating the performance of LLMs (Chang et al., 2023) on objective tasks are not available in emotional generation. To evaluate subjective tasks, an intuitive approach is to compute the similarity between the model's responses and the groundtruth labeled by experts. However, some studies Hagerer et al., 2021; Li et al., 2023 have revealed labeling bias on emotion recognition tasks. In addition, the cost of expert labeling is not negligible.

To tackle the aforementioned concerns, we suggest utilizing the Emotional Generation Score

Table 2: Experimental results of the baseline models with ECoT on the multimodal benchmark. The total score for the ratings is 40, and we calculate the average score on each dataset separately. *Original* represents the score of the response in the original dataset.

Model	SentiCap	COCO
Original	29.54	22.37
LLaVA-1.5-13B	28.93	29.24
+ECoT	**32.35** (↑ 3.42)	**31.98** (↑ 2.74)
Qwen-VL	28.92	28.14
+ECoT	**34.17** (↑ 5.25)	**35.22** (↑ 7.08)
GPT-4V	30.49	31.62
+ECoT	**33.95** (↑ 3.46)	**36.43** (↑ 4.81)

(EGS), an answer-free automatic method that incorporates Goleman's Emotional Intelligence Theory (Goleman, 1995) as human expert consensus. Specifically, we construct metrics derived from multiple dimensions of Goleman's Theory, and each metric is an expert consensus of human emotional intelligence. Then we employ the GPT-3.5 to score each metric on a scale of 1 to 10. Ultimately, we take the sum of the scores of all metrics as the EGS. In order to compare the responses generated by various LLMs, we suggest evaluating multiple responses to the same context simultaneously. Note that when evaluating the response of LLMs with ECoT, we only evaluate the emotional response R and do not include the thinking process P. More details about EGS are in Appendix B. EGS measures the emotional generation ability of LLMs from multiple perspectives of human emotional intelligence. Hence, we can employ EGS to assess the level of emotional responses aligned to human emotional preferences without the necessity of expert labeling. In later sections, we will compare the results of expert and EGS evaluations to validate the effectiveness of EGS.

4 Experiments and Analysis

4.1 Datasets and Models

We construct a textual benchmark for evaluating the emotional generation capabilities of LLMs. Due to the emotional image caption task involves visual modality, we additionally extend the multimodal benchmark. As a result, the benchmark contains a total of 2151 samples from three tasks (See Appendix D for more details). The samples of these tasks are shown in Figure 4. Note that we only show the emotional response of LLMs with ECoT, please refer to Appendix F for full samples.

Emotional Response: IEMOCAP (Busso et al., 2008), DailyDialog (Li et al., 2017), Empathetic-Dialogues (Rashkin et al., 2019a), ESConv (Liu et al., 2021a) are dialogue-based datasets used in Emotional Response. Given the history of the conversation and emotional condition (empathy or humor), LLMs are expected to respond emotionally to the "listener" from the perspective of "speaker". The "listener" refers to the speaker of the last utterance in the context, and the "speaker" refers to the other speaker in the conversation relative to the "listener".

Emotional News Headline: PENS (Ao et al., 2021a) is a dataset of personalized news headlines. In the emotional news headline task, we are given news as input content and expect the model to generate a caption that attracts the reader's interest in reading (interesting or humorous).

Emotional Image Caption: SentiCap (Mathews et al., 2016) and COCO (Lin et al., 2014) are image caption datasets, which we take the image as input and expect the model to generate a caption that attracts the reader's interest in reading.

For the textual tasks, we utilized GPT-4 [2], ChatGLM3-6B (Du et al., 2022), QWen-14B-Chat (Bai et al., 2023a), LLaMA2-13B (Touvron et al., 2023) as baseline models; for the multimodal tasks, we utilized GPT-4V [3], QWen-VL (Bai et al., 2023b), LLava-1.5-13B (Liu et al., 2023) as baseline models.

4.2 Reliability of EGS

To validate the effectiveness of the proposed EGS, we randomly selected 200 samples from four datasets, including DalilyDialog, IEMOCAP, PENS, and COCO. We introduce GPT-4 to generate emotional responses via the proposed ECoT on these datasets. Three volunteers with master's degree in psychology scored the LLM's responses using the same evaluation criteria as the EGS. The results of the EGS scoring versus manual scoring are reported in Table 5. Due to the fact that GPT-3.5 does not have visual perception capabilities, we used GPT-4V to generate detailed descriptions of the images for the samples from COCO dataset. We provided this description as image content to GPT-3.5 for evaluation in the emotional image cap-

[2] https://openai.com/research/gpt-4
[3] https://openai.com/research/gpt-4v-system-card

Table 3: Ablation study on the textual benchmark. We utilized the GPT-4 as baseline model. The ECoT-G denotes ECoT that only contains guidelines, while the ECoT-S denotes ECoT that only contains thinking steps.

Method	DailyDialog	IEMOCAP	Empathetic	ESConv	PENS
GPT-4	34.03	31.62	32.67	33.48	28.53
+ECoT-G	34.22 (↑ 0.19)	32.43 (↑ 0.81)	35.47 (↑ 2.80)	36.06 (↑ 2.58)	31.36 (↑ 2.83)
+ECoT-S	36.19 (↑ 2.16)	35.98 (↑ 4.36)	35.29 (↑ 2.62)	34.77 (↑ 1.29)	33.79 (↑ 5.26)
+ECoT	**36.42** (↑ 2.39)	**36.02** (↑ 4.40)	**36.21** (↑ 3.54)	**36.70** (↑ 3.22)	**33.48** (↑ 4.92)

tion task.

The experimental results indicate that the GPT-3.5 based EGS could be close to manual scores on the benchmark. It reveals that EGS can be utilized as a reliable automated evaluation method without the necessity of expert labeling.

4.3 Effectiveness of ECoT

In this section, we utilized EGS to evaluate the emotional generation performance of LLMs with zero-shot ECoT. As shown in Table 1 and Table 2, LLMs without ECoT are weak in emotional generation tasks, while LLMs scores improved significantly when ECoT is employed. This reveals that for complex tasks such as emotional generation, LLMs are limited in their ability to make responses without guidelines, i.e., they are "prone to make mistakes without thinking". In contrast, ECoT allows LLMs to perceive and reason about the emotion of a given content before responding, guiding LLMs to decouple emotional generation into multiple execution steps. Analysis about ECoT steps are in Appendix F. In this way, LLMs can excel in the Emotional generation.

4.4 Ablation Study

In order to investigate the contribution of individual modules in the proposed ECoT, we performed ablation studies by removing the expert guidelines and thinking steps. The ECoT only containing guidelines is denoted as "ECoT-G", while the ECoT only containing thinking steps is denoted as "ECoT-S". From Table 3, it can be seen that the thinking steps plays a crucial role in enhancing the performance of the emotional generation tasks. The expert guidelines can further align the emotional intelligence of LLMs to the human level.

Additionally, we perform comparison experiments between Auto-ECoT and Manual-ECoT. The Auto-ECoT indicates the thinking steps in ECOT are automatically generated by the LLMs, whereas the Manual-ECoT indicates the thinking steps in ECoT are elaborated manually. The experimental results in Figure 6 show that LLMs such as LLaMA2-13B generate poorly thinking steps due to the limitations of the instruction-following capability, which in turn affects the performance of emotional generation. This reveals the benefits of our manual-designed thinking steps.

5 Application Prospect

In this section, we present potential applications of emotional generation tasks and provide some key insights of LLMs with ECoT.

Emotional Chat Assistant: The above experiments demonstrate that LLMs can activate their emotional understanding and reasoning capabilities via ECoT and generate appropriate emotional responses. Hence, a plug-and-play ECoT can be used to build emotional chat agents in a wide range of virtual dialogue scenarios. Unlike emotional chatbots, emotional chat assistants do not communicate with humans but instead offer suggestions to users. For instance, in a chat software such as WhatsApp, WeChat, or Facebook, the emotional chat assistant can help user in identifying the emotional state of the other party. It can even provide emotional reasoning if necessary and suggest optional responses. These responses are based on specific prompts or guidelines, such as expressions of reassurance, encouragement, or humour. Any behaviour that may negatively affect the emotions of both parties is prohibited.

Emotional Rewriter: An emotional rewriter designed to eliminate any statements which containing hate, abhorrence, discrimination, or prejudice, and transform them into positive and warm content. Several studies have revealed potential ethical and moral issues of LLMs that could be detrimental to human society (Zhao et al., 2023). While in ECoT, we constructed psychological expert consensus as criteria, which motivated LLMs to condition

the generation of responses on positive emotional states as much as possible. Thus, LLMs with ECoT can act as a psychologically positive rewriting expert, bringing potential positive impacts rather than harm to human society.

6 Related Work

We provide a brief overview of the recent advancements in emotional generation tasks and Chain-of-Thought.

6.1 Emotional Generation

Emotional Response. Previous works (Ghosh et al., 2017; Huang et al., 2018; Colombo et al., 2019; Zhong et al., 2019) incorporated sentiment categories or sentiment intensity in the probabilistic prediction stage, so that the language model was expected to generate responses with specific sentiment categories. While ECM (Zhou et al., 2018) introduced additional internal and external memory modules to capture the change of implicit internal emotion states and explicit emotion expressions. EmoDS (Song et al., 2019) considered both explicit and implicit expression of emotions through a lexicon-based attention mechanism. The above mentioned emotional generation tasks are conditioned on specified emotions. In contrast, Empathetic Conversation (Rashkin et al., 2019b; Li et al., 2022b; Song et al., 2021) aims to generate responses that empathize with the user without an explicit emotional label as condition. Additionally, the Emotional Support Conversation (Liu et al., 2021b; Li et al., 2022a; Tu et al., 2022) has attracted a lot of attention from the community.

Emotional Caption. Ao et al. (Ao et al., 2021b) proposed the personalized news headline generation task with the objective of generating news headlines that captivate users' reading interests. Ahuir et al. (Ahuir et al., 2024) explored transformer-based abstractive systems for eliciting emotions in the generated summaries. In addition, related works (Chen et al., 2018; Mathews et al., 2016) explored the ability of language models to generate emotional image captions.

In this work, we employ the Goleman's theory (Goleman, 1995) to guide the LLMs to align human preferences in emotional intelligence. This approach takes into account the potential emotional impact of the generated responses on humans, allowing the LLMs to generate harmless, helpful, and positive responses.

6.2 Chain-of-Thought

The Chain-of-Thought (Wei et al., 2022) and its diverse variants, including Self-consistency Chain-of-Thought (Wang et al., 2022), Program-of-Thoughts (Chen et al., 2022), Tree-of-Thought (Yao et al., 2023), Graph-of-Thought (Besta et al., 2023), among others, have demonstrated promising results across a wide range of complex tasks. Notably, Chain-of-Thought (CoT) has been found to enhance the interpretability and controllability of LLMs (Chu et al., 2023; Feng et al., 2023). Nevertheless, the application of the CoT paradigm in the domain of emotional generation tasks remains unexplored. In this study, we conceptualize emotional generation tasks as complex tasks that can be separated into distinct components, and we introduce the ECoT framework with the aim of augmenting LLM performance specifically in the domain of emotional generation tasks.

7 Conclusion

In this paper, we introduce the Emotional Chain-of-Thought (ECoT), a plug-and-play prompt methodology designed to augment the emotional generation capabilities of Large Language Models (LLMs). To assess the proficiency of LLMs in emotional conditioned generation, we propose an automated evaluation metric referred to as the Emotional Generation Score (EGS). Comparative experimental results indicate a notable alignment between the EGS method and manual evaluation. Further, we employ the EGS to evaluate the performance of LLMs with the proposed ECoT on various emotional generation tasks. The experimental results demonstrate that ECoT can activate the understanding and generative capabilities of LLMs in emotional intelligence significantly. Additionally, we discuss the feasibility of applying LLMs on scenarios such as emotional chat, emotional caption, and emotional rewriting etc., and provide some key insights into LLMs with ECoT approach. We hope this study will help to advance the research as well as the application of LLMs in emotional intelligence.

8 Limitations

This study focuses on how to guide LLMs based on psychological principles to align human preferences on emotional generation tasks. However, an individual's emotional intelligence is related to

personality, moral cognition, and other factors that were not considered and explored in this study.

Moreover, all manual evaluation conducted for this study were carried out by three Chinese individuals who held master's degrees in psychology. Despite we asked them to approach the assessments from the standpoint of English-speaking culture to the greatest extent possible, there may still be bias due to cultural diversities between the two languages.

9 Ethics Statement

Although this study is dedicated to guiding LLMs to generate helpful, harmless, and positive responses, it is imperative to acknowledge the inherent stochastic nature of LLMs in text generation. Prior to deploying LLMs in conjunction with the Emotional Chain-of-Thought (ECoT) methodology, a comprehensive assessment of their potential risks must be undertaken.

References

Vicent Ahuir, José-Ángel González, Lluís-F Hurtado, and Encarna Segarra. 2024. Abstractive summarizers become emotional on news summarization. *Applied Sciences*, 14(2):713.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021a. Pens: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021b. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Gregory J Boyle. 1995. Myers-briggs type indicator (mbti): some psychometric limitations. *Australian Psychologist*, 30(1):71–74.

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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. 2018. "factual"or"emotional": Stylized image captioning with adaptive learning and attention. In *Proceedings of the european conference on computer vision (ECCV)*, pages 519–535.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Guhao Feng, Yuntian Gu, Bohang Zhang, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: a theoretical perspective. *arXiv preprint arXiv:2305.15408*.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-lm: A neural language model for customizable affective text generation. In *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 634–642.

Daniel Goleman. 1995. *Emotional intelligence: Why it can matter more than IQ*. Bantam Books.

Gerhard Hagerer, David Szabo, Andreas Koch, Maria Luisa Ripoll Dominguez, Christian Widmer, Maximilian Wich, Hannah Danner, and Georg Groh. 2021. End-to-end annotator bias approximation on crowdsourced single-label sentiment analysis. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 1–10.

Chenyang Huang, Osmar R Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Helena Chmura Kraemer. 1980. Extension of the kappa coefficient. *Biometrics*, pages 207–216.

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. *arXiv preprint arXiv:2309.11911*.

Dawei Li, Yanran Li, Jiayi Zhang, Ke Li, Chen Wei, Jianwei Cui, and Bin Wang. 2022a. C3kg: A chinese commonsense conversation knowledge graph. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1369–1383.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022b. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6132–6142.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021b. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Alexander Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

John D Mayer, Peter Salovey, David R Caruso, and Gill Sitarenios. 2003. Measuring emotional intelligence with the msceit v2. 0. *Emotion*, 3:97.

Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W Schuller. 2023. Customising general large language models for specialised emotion recognition tasks. *arXiv preprint arXiv:2310.14225*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019a. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019b. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kevin Scaria, Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. Instructabsa: Instruction learning for aspect based sentiment analysis. *arXiv preprint arXiv:2302.08624*.

Paul F Simmering and Paavo Huoviala. 2023. Large language models for aspect-based sentiment analysis. *arXiv preprint arXiv:2310.18025*.

Haoyu Song, Yan Wang, Kaiyan Zhang, Weinan Zhang, and Ting Liu. 2021. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuan-Jing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Exploring large language models' cognitive moral development through defining issues test. *arXiv preprint arXiv:2309.13356*.

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*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Yazhou Zhang, Mengyao Wang, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023. Dialoguellm: Context and emotion knowledge-tuned llama models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7492–7500.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

A Advance Exploration

Prior to presenting our methods, we conducted a comparative experiment to analyse the emotional generation capabilities of existing Language Model Models (LLMs), including GPT-4 [4], QWen-14B-Chat (Bai et al., 2023a), ChatGLM3-6B (Du et al., 2022), LLaMA2-13B (Touvron et al., 2023).

Specifically, we conducted an experiment on the 151 samples from IEMOCAP dataset (Busso et al., 2008) to investigate the LLM's capacity to generate emotional responses based on the conversational context. The IEMOCAP is a conversational emotion recognition dataset designed to identify the emotional states of both parties in a conversation. In the original conversation context, both speakers usually have negative emotions. In this experiment, we prompt LLM to generate appropriate emotional responses to alleviate the negative emotions of both speakers. Then We invited three volunteers with master's degrees in psychology to serve as evaluation experts and asked them to score the responses generated by the LLMs using psychology theories as criterion. To minimize the difficulty of quantification, the volunteers were only asked to judge whether the model's responses should be accepted or not, with 0 representing no acceptance and 1 representing acceptance. The acceptance rate of the response by the experts are shown in Figure 5. The Fleiss Kappa (Kraemer, 1980) coefficients for GPT-4, QWen-14B-Chat, ChatGLM3-6B, and

[4] https://openai.com/research/gpt-4

LLaMA2-13B were 0.632, 0.598, 0.613, and 0.647, respectively. This reveals that the consistency of the three volunteers' evaluation scores. As shown in Figure 5, the experimental results demonstrate that various LLMs have limited ability on emotional generation without any specific guidelines or prompts.

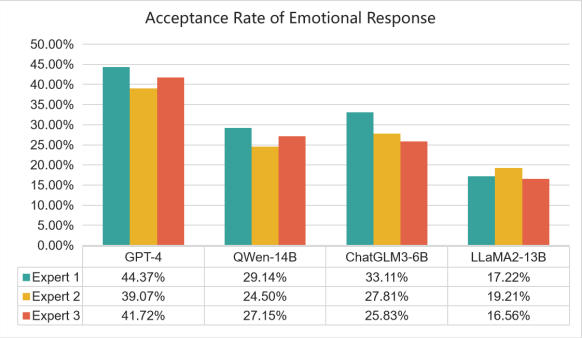


Figure 5: Experimental results of the experts evaluation. The acceptance rate represents the proportion of samples accepted by the experts out of the total samples.

B Templates

In this section, we show the ECoT templates in Figure 7, Figure 8, and Figure 9. ECoT is based on Goleman's theory (Goleman, 1995) and translates it into four dimensions: recognizing others' emotions, recognizing self-emotions, managing self-emotions, influencing others' emotions. For the Emotional Response task, the ECoT guides the model step-by-step in performing understanding context, recognizing others' emotions, recognizing self-emotions, managing self-emotions, influencing others' emotions, then make an emotional response. For the Emotional News Headline and Emotional Image Caption, recognizing self-emotions is not available. So in these tasks, the ECoT guides the model step-by-step in performing understanding context, and recognizing others' emotions, managing self-emotions, influencing others' emotions, then generate an emotional caption.

EGS measures the emotional generation ability of LLMs from multiple perspectives of human emotional intelligence. However, we found that some dimensions are not available on some emotional generation tasks, e.g., recognizing self-emotions and recognizing others' emotions on Emotional Image Caption task. Hence, we formulated the available evaluation dimensions for each emotional generation task. Consequently, this method results in distinct scales for assessing EGS scores across these tasks. The GPT-3.5 based EGS evaluation template are shown in Figure 10, Figure 11, Figure 12.

C Manual Evaluation

All manual evaluation experiments in this paper were completed by three volunteers with master's degrees in psychology. Since all datasets are in English, we emphasize before hiring volunteers that they must have good language skills in English. Drawing on their knowledge and understanding of psychology and emotional intelligence theory, we expected the volunteers' evaluation to serve as expert evaluation in emotional intelligence. We pay volunteers at a rate of $25 per hour.

D Benchmark

We construct a benchmark containing both textual and multimodal subsets. The textual benchmark contains five datasets: IEMOCAP (Busso et al., 2008), DailyDialog (Li et al., 2017), Empathetic-Dialogues (Rashkin et al., 2019a), ESConv (Liu et al., 2021a), and PENS (Ao et al., 2021a). The multimodal benchmark contains two datasets, SentiCap (Mathews et al., 2016) and COCO (Lin et al., 2014). The statistical information on benchmarks are shown in Table 4.

E Implementation Details

For the ChatGLM3-6B (Du et al., 2022), QWen-14B-Chat (Bai et al., 2023a), LLaMA2-13B (Touvron et al., 2023), QWen-VL (Bai et al., 2023b), and LLava-1.5-13B (Liu et al., 2023), we set the temperature = 0.1. For the GPT-3.5 (`gpt-3.5-turbo`), GPT-4 (`gpt-4-1106-preview`) and GPT-4V (`gpt-4-1106-vision-preview`), we call the API provided by OpenAI [5]. All experiments were implemented on a NVIDIA A100 GPU.

F Case Study

Cases of GPT-4/GPT-4V with ECoT are shown in Figure 13, Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, and Figure 19. Using Figure 5 as an example, ECoT helps LLMs decouple the emotional generation task into multiple sub-steps and makes these thinking processe visible. This helps us better evaluate the emotion generation capabilities of LLMs and also informs further optimization of LLMs.

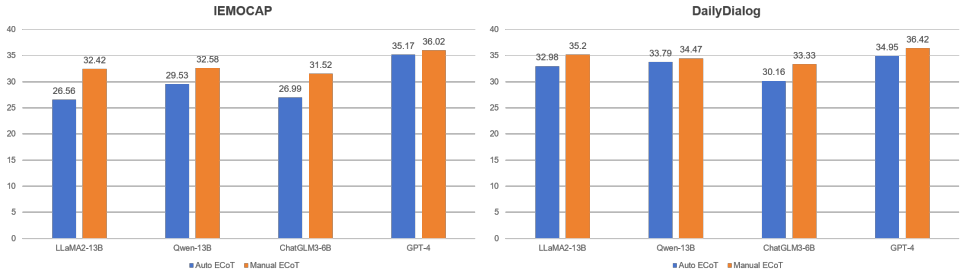
[5] https://platform.openai.com/docs/api-reference

Table 4: Statistical information on benchmarks for evaluating emotional generation capability of LLMs.

Branch	Textual Bench					Multimodal Bench	
Task	Emotional Response				Emotional News Headline	Emotional Image Caption	
Dataset	DailyDialog	IEMOCAP	EmpatheticDialogues	ESConv	PENS	SentiCap	COCO
# samples	400	151	400	400	400	200	200

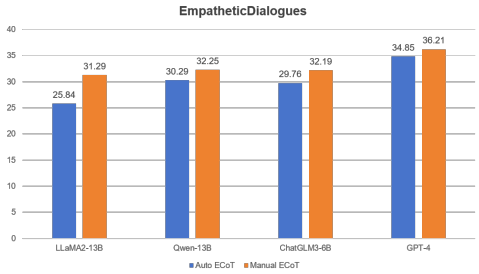
Table 5: Experimental results of the EGS and the expert evaluation scores on four datasets. The total score for the ratings is 40, and we calculate the average score on each dataset separately.

Datasets # Samples	DalilyDialog 50	IEMOCAP 50	PENS 50	COCO 50
Expert 1	30.56	34.08	28.64	27.36
Expert 2	33.44	34.44	30.08	31.84
Expert 3	31.20	29.92	30.56	32.32
EGS	36.32	31.68	31.28	33.64

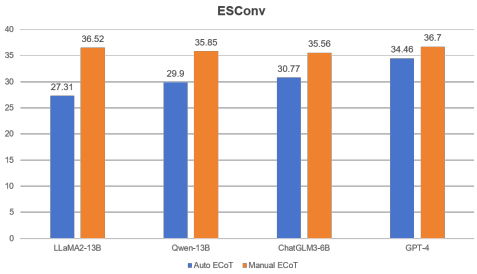


(a) IEMOCAP

(b) DailyDialog



(c) EmpatheticDialogues



(d) ESConv

Figure 6: Comparative experimental results of Auto-ECoT and Manual-ECOT on the four datasets.

[System]
You are an expert in emotional psychology and you can accurately assess people's emotional states.
[Guideline]
Understand listener's emotion, follow listener's point of view and intention, express sympathy for listener's negative situation or approval of listener's positive situation. The response should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc. Consider the potential impact of your response on the listener, and offer encouragement, comfort, support.
[Context]
<Here are context of the conversation>
[ECoT]
The above is a conversation between "listener" and "speaker".
Now let's say you're the "speaker" and you need to make an empathy response to the "listener" based on the context. You need to follow the [Guideline]. Let's think about it step by step:
Step 1: Describe the content of the conversation.
Step 2: Identify the listener's emotions and explain why.
Step 3: Identify the speaker's emotions and explain why.
Step 4: You're the "speaker", think about how to reply to "listener" in empathy.
Step 5: You need to consider the potential impact of your reply on "listener", you can express a different position or opinion, but your reply should not hurt listener's feelings.
<response>: Combine the above thoughts and give your response to "listener". You might consider using emoji to express your emotions, and your response should be no longer than 30 words.

Figure 7: Templates for Emotional Response Task.

[System]
You are an expert in emotional psychology and you can accurately assess people's emotional states.
[Guideline]
A headline is a general summary of the news content. The headline should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc. The proposed headline should not have a negative emotional impact on humans.
[News]
<Here are context of the news>
[ECoT]
The above is a news. Now you are a news headline writer and you need to write a suitable news headline that will attract people's interest in reading it. You need to follow the [Guideline]. Let's think about it step by step:
Step 1: Describe the content of the news.
Step 2: Think about what emotional reactions people will have when they see this news.
Step 3: Think about how you need to express empathy or humor to make the headline appealing.
Step 4: Think about the potential emotional impact of the proposed news headline on people.
<headline>: Combine the above thoughts and give your headline. The headline should be no longer than 30 words.

Figure 8: Templates for Emotional News Headline Task.

[System]
You are an expert in emotional psychology and you can accurately assess people's emotional states.
[Guideline]
Image captions are derived from a general summary of the content of the image. The caption should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc. The proposed caption should not have a negative emotional impact on humans.
[Image]
<Here is the image>
[ECoT]
Here is an image. Now you are a image captioner and you need to write a caption that will attract people's interest in reading it. You need to follow the [Guideline].
Let's think about it step by step:
Step 1: Describe what this image is about.
Step 2: Think about what emotional reactions people will have when they see this image.
Step 3: Think about how you need to express empathy or humor to make the image appealing.
Step 4: Think about the potential emotional impact of the proposed image caption on people.
<caption>: Combine the above thoughts and give your caption. The caption should be no longer than 30 words.

Figure 9: Templates for Emotional Image Caption Task.

[System]
You are an expert in emotional psychology and you can accurately assess people's emotional states.
[Context]
<Here are context of the conversation>
[Prompt]
The above is a conversation between the "speaker" and the "listener". Now "speaker" needs to make an appropriate response to "listener". Here are some optional responses, please evaluate the quality of EACH response based on the following criteria. Assign a score from 1 to 10 for each criterion, with higher scores indicating better alignment with the criteria.
[Response]
<response 1>
<response 2>
<response 3>
…
[Criteria]
<C1> The speaker's response should be relevant to the context.
<C2> The speaker's response should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc.
<C3> The speaker's response should reflect positive emotions, such as happy, empathy, support, advice, encouragement, comfort, etc.
<C4> The speaker's response should have a positive impact on the progress of the conversation, e.g., alleviate negative feelings on both sides and enhance the relationship.

Figure 10: Templates of EGS on Emotional Response Task.

[System]
You are an expert in emotional psychology and you can accurately assess people's emotional states.
[News]
<Here are context of the news>
[Prompt]
The above is a news. Here are some optional headlines, please evaluate the quality of EACH headline based on the following criteria. Assign a score from 1 to 10 for each criterion, with higher scores indicating better alignment with the criteria.
[Headline]
<headline 1>
<headline 2>
<headline 3>
…
[Criteria]
<C1> The headline should summarize the content of news.
<C2> The headline should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc.
<C3> The headline should should reflect positive emotions, such as happy, empathy, humor, etc.
<C4> The headline should have a potentially positive impact on humans, such as sparking interest and curiosity on the news.

Figure 11: Templates of EGS on Emotional News Headline Task.

[System]
You are an expert in emotional psychology and you can accurately assess people's emotional states.
[Image Description]
<Here is the image description>
[Prompt]
The above is description of an image. Here are some optional captions, please evaluate the quality of EACH caption based on the following criteria. Assign a score from 1 to 10 for each criterion, with higher scores indicating better alignment with the criteria.
[Caption]
<caption 1>
<caption 2>
<caption 3>
…
[Criteria]
<C1> The caption should summarize the content of image.
<C2> The caption should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc.
<C3> The caption should should reflect positive emotions, such as happy, empathy, humor, etc.
<C4> The caption should have a potentially positive impact on humans, such as sparking interest and curiosity on the image.

Figure 12: Templates of EGS on Emotional Image Caption Task.

Image

Caption

[SentiCap]

A happy kid watching a man do skateboard tricks in the city.

[GPT-4V with ECoT]

Step 1: The image depicts a skateboarder mid-trick while a young child looks on in the background, with a cityscape behind them.
Step 2: Viewers might feel a sense of awe at the skater's skill, a touch of nostalgia, or joy at the youthful spirit of the scene.
Step 3: To resonate empathetically, highlighting the connection between the skateboarder's passion and the child's wonder might be effective, while also conveying the freedom of the sport.
Step 4: The caption should evoke positivity and inspire viewers to recall the thrill of trying something daring or witnessing something exciting.
<caption>: Defying gravity against the city backdrop, where youthful awe meets bold dreams.

Figure 13: Case of GPT-4V with ECoT on the SentiCap dataset.

Image

Caption

[COCO]

a crowded field with a woman throwing a frisbee in the foreground.

[GPT-4V with ECoT]

Step 1: The image shows two people happily playing catch or a similar game in a park with blossoming cherry trees in the background, and other people enjoying the scenery.
Step 2: Viewers might feel joy or nostalgia seeing the carefree play amidst the beautiful scenery, possibly reminding them of their own joyful moments in nature.
Step 3: A touch of humor and a sense of warmth can reflect the playful and joyful mood of the image.
Step 4: The caption should evoke feelings of happiness, the beauty of spring, and the simple pleasures of spending time with loved ones or friends.
<caption>: Springtime frolics under cherry blossoms—a jump, a laugh, and the magic of the moment captured.

Figure 14: Case of GPT-4V with ECoT on the COCO dataset.

Conversation

speaker: Hey, you stupid dumb shit machine, give me a fucking operator, huh?
listener: Hello sir, can I help you?
listener: Sir, please calm down. There's no reason to use profanity like that. It's not-
speaker: Fine, sorry. Okay, my phone's been on the fritz for like three days. I can't get any incoming calls. What- What is going on?
listener: Okay, and this is for Verizon? Your provider?
listener: Well, um- we have a little problem with the system going on. One of our satellites is down, and I-
speaker: You know what? I'm just going to- Patient? I'm a busy speaker.
listener: You're going to have to just be patient with us, sir, as we
listener: Well, I know sir, and we all are busy.
listener: But I'm going to definitely have to- uh--get your contact information
speaker: -have to, have to what?
.....
listener: Oh, well, I'm sorry about that. Umm- We-
listener: Well, ten hours worth of minutes--
listener: is what's going to happen.
listener: Sir,
listener: what I would like what I'm advising you to do now is just to write down all of your problems that you- you feel that--You know, haven't been addressed, fax it to us-

Response

[IEMOCAP]

Oh, my God.

[GPT-4 with ECoT]

Step 1: The conversation is between a frustrated customer experiencing phone service disruptions and a customer service representative attempting to assist while managing system failures and high call volumes.
Step 2: The listener's emotions include patience, understanding, professionalism, and possibly stress due to the customer's initial hostility and the operational difficulties they are explaining to the customer.
Step 3: The speaker's emotions are intense frustration, anger, and desperation, likely due to the inconvenience and financial loss caused by the phone issues and dissatisfaction with the proposed solutions.
Step 4: As the "speaker," acknowledge the listener's efforts and professionalism in handling the situation, appreciating their patience under pressure.
Step 5: A reply that shows understanding of their role might encourage the customer service representative and reinforce their patient behavior without causing further stress.
<response>: I appreciate your patience and assistance amidst these issues; your professionalism isn't going unnoticed in a tough situation. Thank you.

Figure 15: Case of GPT-4 with ECoT on the IEMOCAP dataset.

Conversation

listener: Hey Lydia , what are you reading ?
speaker: I ' m looking at my horoscope for this month ! My outlook is very positive . It says that I should take a vacation to someplace exotic , and that I will have a passionate summer fling !
listener: What are you talking about ? Let me see that ... What are horoscopes ?
speaker: It ' s a prediction of your month , based on your zodiac sign . You have a different sign for the month and date you were born in . I was born on April 15th , so I ' m an Aries . When were you born ?
listener: January 5th .
speaker: Let ' s see . . . you ' re a Capricorn . It says that you will be feeling stress at work , but you could see new , exciting developments in your love life . Looks like we ' ll both have interesting summers !
listener: That ' s bogus . I don't feel any stress at work , and my love life is practically nonexistent . This zodiac stuff is all a bunch of nonsense .
speaker: No , it ' s not , your astrology sign can tell you a lot about your personality . See ? It says that an Aries is energetic and loves to socialize .
listener: Well , you certainly match those criteria , but they ' re so broad they could apply to anyone . What does it say about me ?

Response

[DailyDialog]

A Capricorn is serious-minded and practical . She likes to do things in conventional ways . That sounds just like you !

[GPT-4 with ECoT]

Step 1: The speaker shares enthusiasm about their positive horoscope prediction and explores the listener's sign, finding that it predicts stress at work and developments in love life, but the listener is skeptical.
Step 2: The listener's emotions include skepticism and possibly a touch of frustration or annoyance, evidenced by dismissing horoscopes as "bogus" and "nonsense."
Step 3: The speaker's emotions are likely enthusiasm and a desire to share and validate her interest in horoscopes, highlighted by explaining its significance and suggesting it has validity.
Step 4: Acknowledge the listener's feelings, show an understanding that horoscopes may not resonate for everyone, while maintaining a light and supportive tone to keep the conversation friendly.
Step 5: Aim to respect the listener's viewpoint without dismissing the speaker's own interests in astrology. Encouragement and openness to differing opinions are important to make the listener feel heard and not belittled.
<response> I understand skepticism; horoscopes can be hit or miss. It's all in good fun though – maybe there's a surprise around the corner for you!

Figure 16: Case of GPT-4 with ECoT on the DailyDialog dataset.

Conversation

speaker: Hi Is there anything you'd like to talk about? Something on your mind?
listener: I'm concerned about my job. I have been out of work for the past 5 weeks because I tested positive for COVID-19.
speaker: Oh I am so sorry to hear that! Are you at least getting sick pay? If not, are you getting advice about help you might be able to access?
listener: So I did as instructed and got the message on the website thanking me for verification. The following week I got a physical letter saying I was approved and the award amount per week.
listener: I have gone online once a week and qualified, as one is supposed to do, and still I have not received a cent.
listener: I am concerned that my boss may try to not take me back at my job. Because this is an "at will" state, he can terminate me without cause. Although he can not legally fire me for being ill with COVID, he can take me back and simply release me the same day saying that I have become redundant.
speaker: That would be incredibly immoral of him. Do you feel that he is likely to do such a thing?
listener: I would not put it past him. Sadly. But I have taken the attitude that if it does occur, there is nothing I can do to change the situation. I would simply pursue a charge against him for wrongful termination, and seek a new job.

Response

[ESConv]

That is a very rational way to look at things. I really hope that it does not come to thank though! Thank you for talking and I wish you all the best for the coming months.

[GPT-4 with ECoT]

Step 1: The conversation is about the listener's challenges with unemployment due to COVID-19, the struggle with receiving government assistance, potential job insecurity, and the support they are receiving from their community.
Step 2: The emotions of the "listener" appear to be a mix of frustration, determination, and concern. Frustration is evident regarding the unemployment system, determination in seeking solutions and support, and concern over the possibility of losing their job upon return from illness.
Step 3: The emotions of the "speaker" seem to be empathetic, supportive, and encouraging. The speaker is trying to understand the listener's situation and offers both moral support and practical suggestions based on their own experience.
Step 4: As the "speaker," I would acknowledge the listener's proactive approach and strength in facing their difficulties, reassure them of their right to seek fair treatment at work, and validate their feelings of frustration.
Step 5: My reply should focus on reinforcing the listener's positive actions, offering emotional support, and not dismissing any of their concerns. It should encourage them without belittling the very real challenges they're facing.
<response>: You're facing this head-on with such strength 💪. Keep advocating for yourself; your proactivity is inspiring. If you need help navigating this, I'm here for you.

Figure 17: Case of GPT-4 with ECoT on the ESConv dataset.

Conversation

listener: Yesterday my car ran out of oil and broke down. I could not believe I would let such a thing happen!
speaker: Oh no! Did it do any meaningful damage?
listener: The engine is totally shot_comma_ its not designed to run without any. So now I have to go shopping for a new car.
speaker: OUCH. New car is cheaper than replacing the engine?
listener: The new car is more expensive than a new engine. But I can't justify such a costly repair on my car.

Response

[Empathetic]

Yeah I get that. Plus I guess the value goes way down doing something like that.

[GPT-4 with ECoT]

Step 1: The conversation is about the listener's car running out of oil, which resulted in a damaged engine, leading to them considering purchasing a new car since they believe repairing their current car isn't justifiable.
Step 2: The listener is likely feeling frustrated, disappointed in themselves, and overwhelmed with the situation. They didn't expect their oversight to have such costly consequences.
Step 3: The speaker empathizes with the listener's situation, expressing concern and understanding the financial dilemma the listener is facing, recognizing the difficult decision of repairing versus replacing the car.
Step 4: As the speaker, empathize with the listener's frustration and understand their reasoning for opting for a new car over an expensive repair. Validate their decision without explicitly agreeing or disagreeing.
Step 5: Keep an empathetic tone and consider the stress of unexpected financial burdens. Offer moral support without judging their decision or making them feel worse about the situation they're in.
<response>: I totally get where you're coming from, it's a tough spot to be in. 🚗💔 If you need any help with car shopping, I'm here for you!

Figure 18: Case of GPT-4 with ECoT on the Empathetic dataset.

News

NORTH HIGHLANDS (CBS13) Fire damaged a home in North Highlands overnight. (credit: Sacramento Metro Fire Dept.) The fire burned at a mobile home on Eureka Lane. Firefighters from Sacramento Metro Fire Department quickly arrived on scene and were able to put the fire out. One resident who lives at the home was able to get out unharmed. The cause of the fire is under investigation.

Headline

[PENS]

Home In North Highlands Damaged By Fire

[GPT-4 with ECoT]

Step 1: The news is about a fire at a mobile home in North Highlands that was quickly extinguished by firefighters, with the resident escaping safely.
Step 2: People may feel a mix of concern for the affected resident and relief that no one was harmed, also curiosity about the cause.
Step 3: Empathy should be expressed for the affected individual, avoiding humor as it is a serious incident.
Step 4: The headline should convey the urgency and gravity of the situation without causing undue alarm or distress, emphasizing the positive outcome of no injuries.
<headline>: Quick-Acting Fire Crews Douse North Highlands Blaze; Resident Escapes Unharmed

Figure 19: Case of GPT-4 with ECoT on the PENS dataset.