CAPE: A Chinese Dataset for Appraisal-based Emotional Generation using Large Language Models

June M. Liu[1,2*†] **He Cao**[1,3*†] **Renliang Sun**[1,4*] **Rui Wang**[1] **Yu Li**[1] **Jiaxing Zhang**[1]
[1] International Digital Economy Academy (IDEA), [2] University of Michigan, Ann Arbor
[3] Hong Kong University of Science and Technology [4] University of California, Los Angeles
juneliu@umich.edu, hcaoaf@connect.ust.hk, sunrenliang@ucla.edu,
{wangrui,liyu,zhangjiaxing}@idea.edu.cn

Abstract

Generating emotionally appropriate responses in conversations with large language models presents a significant challenge due to the complexities of human emotions and cognitive processes, which remain largely underexplored in their critical role in social interactions. In this study, we introduce a two-stage automatic data generation framework to create CAPE, a Chinese dataset named **C**ognitive **Ap**praisal theory-based **E**motional corpus. This corpus facilitates the generation of dialogues with contextually appropriate emotional responses by accounting for diverse personal and situational factors. We propose two tasks utilizing this dataset: emotion prediction and next utterance prediction. Both automated and human evaluations demonstrate that agents trained on our dataset can deliver responses that are more aligned with human emotional expressions. Our study shows the potential for advancing emotional expression in conversational agents, paving the way for more nuanced and meaningful human-computer interactions.

Figure 1: An example of an emotion generation conversation with the analysis on the results of LLMs.

1 Introduction

Emotion is a crucial aspect of human-computer interaction (Brave and Nass, 2007), especially with large language models (LLMs) (Ratican and Hutson, 2023; Sabour et al., 2024), as generating suitable emotional responses is essential for making communications natural and machines more human-like (Hortensius et al., 2018; Li et al., 2022; Kang et al., 2024; Sun et al., 2024). Expressing human-like emotions is challenging for machines, as it involves a complex psychological process that requires considering personal traits, situational influences, and an individual's evaluation of the current scenario(Masters, 1991; Greenaway et al., 2018; Winberg et al., 2014). Furthermore, emotion does not exist as a single entity but is generated as a collection of subjective experience and behavior (Gross and Feldman Barrett, 2011).

Existing LLMs may fall short of capturing the nuances of human emotions, making their interactions seem impersonal or inadequate. For instance, in Figure 1, we present an example conversation that showcased each speaker's background and utterance. As we can see, GPT-4 (OpenAI et al., 2024) and Llama3-8B-Chinese (Wang et al., 2024) might generate responses that, although contextually relevant, fail to align with the emotional tone required in the given setting. This inadequacy stems from a lack of depth in understanding and simulating the complex process of human emotion generation. Furthermore, existing studies have primarily concentrated on English, despite the pivotal role of language in conveying emotions (Lindquist, 2017). Research on emotional expression in other

∗ Equal contribution. † Work is done during their internship at IDEA.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6306–6324

languages and cultures has received limited attention in comparison.

To address this gap in emotional understanding, we turn to Cognitive Appraisal Theory (CAT) (Lazarus, 1991), a psychological framework that comprehensively explains how emotions are generated through the appraisal of external stimuli. At the core of CAT is the concept of *appraisal*, which describes how individuals evaluate external stimuli, thereby eliciting emotional responses. This process underscores the dynamic role of interpretation in shaping emotions—such as when encountering an animal that a person believes may hurt him or her, the person might immediately appraise it as a threat, triggering fear.

Building on CAT, we introduce **CAT-BEAR**: a **C**ognitive **A**ppraisal **T**heory **Ba**sed **E**motionally **A**ppropriate **R**esponse framework designed to enhance LLMs' ability to express emotions accurately. CAT-BEAR comprises three main components: *1) intra-individual factors*, including personal factors and situational construal; *2) the appraisal process*, which involves evaluating external stimuli based on these factors and the conversation history; and *3) appraisal outcomes*, which are the resulting emotions and action tendencies. This structured approach ensures broad applicability across individuals with diverse intra-individual factors and different contexts, integrating emotions and actions into the ongoing conversation history. Given the importance of cultural context in emotional expression, our study specifically focuses on the use of Chinese, allowing us to tailor the dialogues to align with the nuances of Chinese language and culture.

Using the CAT-BEAR framework, we generate a dataset of multi-turn dialogues between two individuals, each assigned a unique personality, goal, and situational construal. GPT-4-turbo generates relevant beliefs and knowledge to enrich each person's profile. Using this background, it sequentially produces individual emotion labels and utterances according to appraisal process guidelines. This automatic data synthetic framework yields **CAPE** (**C**ognitive **Ap**praisal theory-based **E**motional corpus), a dataset of 2,848 multi-turn dialogues covering 15 distinct emotions. The raw dataset undergoes a thorough cleaning and rigorous human evaluation. We utilize evaluation metrics from related studies—label correctness, emotion-utterance alignment, emotion-context alignment, intensity, coherence, and fluency—and enlist three raters to assess data quality. This process ensures accuracy in emotion labeling, contextual coherence, and overall conversational fluency.

The practical use of dataset CAPE is shown through the creation of a fine-tuned model that can engage in new dialogues while effectively expressing human-like emotions. We test the model's performance through two tasks: predicting the next speaker's emotion label and their utterance in a dialogue. For the first task, we evaluate how well the predicted emotion labels match the actual labels. Since traditional metrics, like exact-match, treat emotions as separate categories and ignore their nuanced nature, we include emotional distance as an additional metric to capture subtle similarities more effectively. In the second task, we compare how closely the utterances generated by our model align with the ground truth. Through these two tasks, we show that our dataset enables the model to effectively capture and convey nuanced human emotions in dialogues. In summary, our study offers the following contributions:

- We propose CAT-BEAR, an automatic data generation framework based on cognitive appraisal theory that addresses the challenge of aligning generated emotions with human-like emotions in dialogues.
- We then construct an emotional dialogue dataset, CAPE, with rigorous quality control and human evaluation to ensure accurate emotion labels and contextually appropriate utterances.
- We design two evaluation tasks: predicting the next speaker's emotion label and generating an appropriate emotional response. Our model fine-tuned on CAPE significantly outperforms state-of-the-art models in both tasks. We believe our framework advances research on CAT-theory-driven dialog systems for more human-like emotional responses.

2 Related works

Most related studies focus on three topics: textual emotion recognition in conversation, emotional support conversation, and emotion generation.

2.1 Textual Emotion Recognition

Textual emotion recognition (TER) involves the automatic identification of emotions within text and has emerged as a key area of focus in natural language processing due to its significant academic and commercial implications. TER requires in-depth analysis and ongoing optimization of methodologies for accurate emotion prediction (Deng and Ren, 2021). Moreover, the complexity of this topic grows from sentence level to document level, as emotions may be conveyed through subtle meanings, metaphors, sarcasm, and irony (Alswaidan and Menai, 2020). Researchers have applied LLMs to recognize emotions from online posts (Liu et al., 2024b; Yang et al., 2023), TV series (Zhang et al., 2024b; Peng et al., 2024), daily conversations (Fu, 2024), and domain-specific dialogues (Xing, 2024). To improve the accuracy of emotion recognition, many studies have begun to utilize multi-modal inputs, such as text, audio, and video (Zhang et al., 2023; Lei et al., 2023; Gan et al., 2023; Cheng et al., 2024). Despite these advances, there is still room to explore how LLMs can authentically convey emotions for more human-like expressions. Our work addresses this challenge by using cognitive appraisal theory for emotion generation, providing a novel approach to enhancing human-LLM interactions.

2.2 Emotional Support Conversation

Emotional Support Conversation (ESC) aims to offer support to individuals dealing with emotional concerns through social interactions, emphasizing both conversational skills and counseling strategies (Liu et al., 2021). Researchers have employed LLMs to alleviate emotional problems (Zheng et al., 2023) or elicit positive emotions (Zhou et al., 2023b). Recent studies highlight the importance of individual factors in formulating empathetic responses that better align with people's needs. Researchers took help-seekers' persona (Cheng et al., 2022) and situational information (Sabour et al., 2022) into consideration to offer help across different populations. Furthermore, there have been studies analyzing users' cognition and affection (Zhou et al., 2022) or exploring emotion's cause (Yang et al., 2024) to provide more appropriate empathetic responses. Current research on ESC emphasizes formulating empathetic responses but focuses mainly on understanding users' emotional states rather than enabling LLMs to express human-like emotions.

2.3 Emotional Utterance Generation

Studies on emotional utterance generation aim at aligning general human affective responses. Some promising studies adapted related psychological theories to let LLMs generate emotions. For example, Li (Li et al., 2024) proposed ECoT that enhances LLMs' emotional intelligence by incorporating Goleman's Emotional Intelligence Theory (Goleman, 2020). Moreover, Croissant et al. (2024) considered both appraisal and memory systems to simulate affective outputs. While these studies demonstrate the effectiveness of their methods, they face limitations when adapted to real-life conditions across diverse populations: either intra-individual factors are not considered, or the context is confined to gaming scenarios.

3 Method

3.1 Definitions

In Cognitive Appraisal Theory, intra-individual factors shape appraisal outcomes through the appraisal process. These intra-individual factors include situational construal and personal characteristics such as personality, goals, beliefs, and knowledge. Appraisal outcomes, in turn, represent the results generated from this appraisal process. Below, we define each of these concepts in detail:

- **Situational Construal** refers to individuals' representation of situations or their perception of situational variables (Funder, 2016). Unlike the objective situation, situational construal is more subjective and interacts directly with personal factors. To cover as many conditions as possible, we use the 89 different situational construals from the Riverside Situational Q-sort (Funder, 2016). The completed list of all 89 situational construals can be found in Appendix A.1.

- **Personal Factors**. (1) *Personality* are used from the 32 personalities defined in the Big Five personality traits (Roccas et al., 2002). (2) *Goals* are orientations toward specific pursuits and fall into two main types: achievement goals and affiliation goals. Achievement goals focus on success and the effort needed for performance (Elliott and Dweck, 1988), while affiliation goals drive the desire for social contact, prompting individuals to seek emotional support, attention, and positive reinforcement (Hill, 1987). (3) *Beliefs* are what is normatively acceptable, feasible, and legitimate and can lead to various emotions (Lazarus, 1991). It can be divided into empirical (to believe objects or perceived value), relational (to believe someone), and conceptual (to believe in narratives) beliefs (Seitz and Angel, 2020). (4)

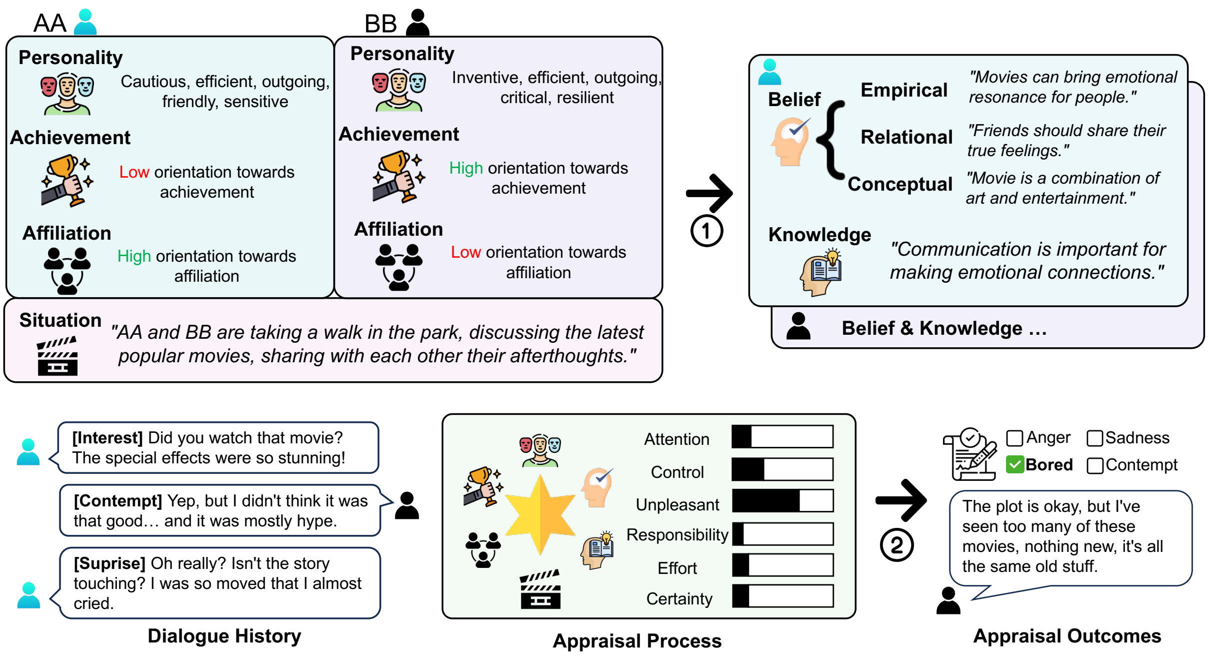

Figure 2: Overview of CAT-BEAR. It contains two stages: (1) AA and BB are initially assigned specific personalities, goals, and situational construal, which are used to generate their beliefs and knowledge (2) the appraisal process, where individuals evaluate the interaction across six dimensions (unpleasantness, control, responsibility, certainty, effort, and attention) to generate emotions and utterances.

Knowledge refers to the understanding of potential harms or benefits (Lazarus, 1991). It can significantly influence how an encounter is appraised as beneficial or not, thereby leading to different emotions.

- **Appraisal Outcomes** includes both emotion and corresponding behavior (i.e., utterances in the dialogue). We employ the 15 emotions from Smith and Ellsworth's study (Smith and Ellsworth, 1985): happiness, sadness, anger, boredom, challenge, hope, fear, interest, contempt, disgust, frustration, surprise, pride, shame, and guilt.

3.2 Data-generative Framework

In this part, we present the data-generative framework CAT-BEAR for simulating dialogue between two individuals. As depicted in Figure 2, the framework unfolds in two stages.

First Stage: generating beliefs and knowledge based on Personality, Goals, and Situation. In this stage, each speaker is assigned a unique personality and set of goals. For example, AA is defined as "cautious, efficient, outgoing, friendly, and sensitive", with low achievement and high affiliation goals. Conversely, BB is "inventive, critical, and resilient", with high achievement and low affiliation goals. The pre-defined situational context—such as "a casual walk in the park and a discussion about popular movies"—is then enriched with these personality traits and goals, providing a more detailed background for the dialogue. Using these attributes, we further prompt GPT-4-turbo-0409 to generate personalized beliefs and knowledge for each individual, as belief and knowledge generation are influenced by an individual's unique characteristics in specific contexts (Lazarus, 1991). The detailed prompt of belief and knowledge generation is listed in Appendix A.2. For instance, AA's belief that "friends should share their true feelings" aligns with a high affiliation goal and the context of reflecting on shared experiences.

Second Stage: simulating appraisal process to generate emotion label and utterance. In this stage, we model the appraisal process, where each speaker assesses their experience based on six dimensions: unpleasantness, control, responsibility, certainty, effort, and attention. By analyzing these dimensions alongside the dialogue history and predefined intra-individual factors, we design a targeted prompt for GPT-4-turbo to predict the speaker's emotion. The model sequentially categorizes the degree of each dimension—beginning with unpleasantness, followed by expected effort, attention, uncertainty, control, and responsibility—to identify the most probable emotion label and the corresponding emotion-aligned utterance. Detailed prompts describing this appraisal process are provided in the Appendix A.3. For example, as illustrated in the bottom right of Figure 2, BB's appraisal, marked by high unpleasantness and low scores on other dimensions, suggests an emotion of *"boredom"* and an utterance expressing disinterest in the movie, aligning with BB's personal factors, situational understanding, and dialogue history.

3.3 Data Quality Control

Manual Dialogue Refinement. To improve data quality, we recruited three native Chinese-speaking workers, all of whom are university students or graduates, to meticulously review and refine dialogues generated by LLMs. Each worker carefully read the instructions and completed trial annotations before beginning the task. The manual filtering process involved: (1) cleaning emotion labels by removing irrelevant ones and aligning labels with the settings, and (2) refining utterances to match emotion labels and modifying phrasing for colloquial language and authentic Chinese conversational tone. For cost reasons, each dialogue was assigned to a single worker, and we randomly selected subsets for quality checks, comparing raw and refined versions.

Data Quality Evaluation. We recruited six additional Chinese-speaking workers, split into two groups, to assess raw and refined data quality. The evaluation metrics, based on prior research (Qian et al., 2023) and tailored to our study, include: (1) *EmoCategory*, assessing whether the emotion label aligns with the settings; (2) *EmoMatch*, evaluating the degree to which the utterance conveys the labeled emotion; (3) *SettingMatch*, measuring if the utterance content aligns with intra-individual factors and situational context; (4) *EmoIntensity*, indicating the intensity of the labeled emotion in the utterance; (5) *Coherence*, determining if the utterance logically fits the conversation context; and (6) *Fluency*, assessing whether the utterance is fluent and easy to understand.

We further curated a human evaluation set by randomly selecting two utterances from each dialogue in the test set, resulting in a total of 283 utterances for rating. Each utterance is provided with a corresponding emotion label, intra-individual factors, and situational construal. We provide each rater with a detailed rating manual outlined in the Appendix A.4. The quality of the human evaluation for the dataset, before and after refinement, is summarized in Table 1. Correlation between raters' scores is significant across dimensions ($p < 0.05$). We observe that manual calibration improves the accuracy of emotion labels and ensures that utterances better align with character settings and dialogue context.

Dimensions	Score Range	Ratings before Filtering	Ratings after Filtering
EmoCategory	0-1	0.89	0.93 (↑4.5%)
EmoMatch	1-5	4.61	4.80 (↑4.0%)
SettingMatch	1-5	4.09	4.52 (↑9.5%)
EmoIntensity	0-2	1.76	1.80 (↑2.3%)
Coherence	1-5	4.93	4.94
Fluency	1-5	4.80	4.85 (↑1.0%)

Table 1: Utterance-level human evaluation on CAPE.

3.4 Dataset Statistics

We named the refined dataset as CAPE (Cognitive Appraisal theory-based Emotional corpus). Table 2 offers a comprehensive overview of the CAPE dataset, encompassing 2,848 dialogues which cover 89 unique situations. The distribution of emotions of CAPE can be found in Figure 3. The average number of utterances per dialogue is around 10, with an average of 38.3 tokens per utterance, indicating a substantial amount of information exchanged within the conversations.

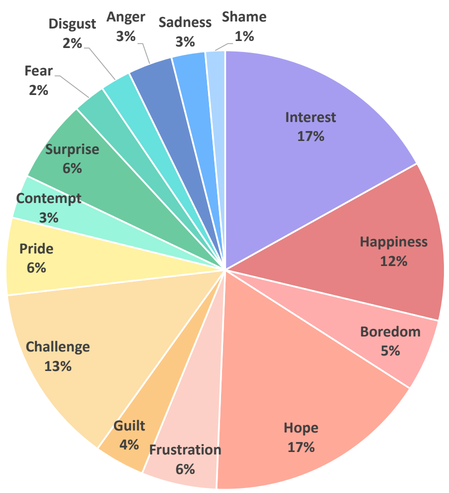

Figure 3: Emotions distribution of CAPE

To evaluate the quality of CAPE, we compare it with representative emotional dialogue datasets, as shown in Appendix A.5. The results indicate that CAPE is uniquely grounded in psychological theories and incorporates a broad range of personal factors closely linked to emotions. It also boasts wider

# Dialogs	2,848
# Utterances	28,643
# Situations	89
Avg. # utterances per dialog	10.0
Avg. # tokens per dialog	385.0
Avg. # tokens per utterance	38.3
# Dialogs in Train Set	2,563
# Dialogs in Validation Set	143
# Dialogs in Test Set	142

Table 2: Statistics of the CAPE dataset.

coverage of emotions and situations compared to existing datasets. Furthermore, unlike datasets annotated from TV series dialogues, CAPE is generated synthetically, which frees it from specific character constraints and limited kinds of storylines, allowing for a more flexible and diverse approach to dialogue generation.

4 Experiments

4.1 Tasks

We have outlined two tasks to evaluate the performance of our dataset, including emotion prediction and appropriate emotional utterances generation.

Task1: Emotion Prediction. This task is to let LLMs predict the emotion label of the next speaker in the dialogue based on intra-individual factors and dialogue history. The predicted emotion label should be chosen within the 15 emotions defined in the CAT-BEAR. It requires LLMs to understand and anticipate the emotions a person may express under specific conditions.

Task2: Next Utterance Prediction. The task asks LLMs to generate the utterance of the next speaker in the dialogue based on intra-individual factors and dialogue history. The quality of the generated utterances reflects a model's ability to express human-like emotional responses.

4.2 Models

Models for Task1. When evaluating the capability of predicting emotions, we compared our model with baselines with high proficiency in Chinese including GLM-4-0520 (GLM et al., 2024), GLM-4-9B-chat (GLM et al., 2024), Llama3-8B-Chinese-Chat (Wang et al., 2024), DeepSeek-v2 (DeepSeek-AI, 2024), Qwen-2.5-72B-Instruct (Team, 2024), and GPT-4-0613 (OpenAI et al., 2024). For all LLMs, we report zero-shot and 4-shot results.

Models for Task2. When examining the generation of appropriate emotional utterances, we incorporated two expert models alongside the LLMs mentioned in the emotion prediction task. The first is CharacterGLM-3 (Zhou et al., 2023a) which supports role-playing based on character settings, and the second one is Emohaa (Sabour et al., 2023) which is designed to provide empathetic support. Following the specified input requirements for inference, we report only zero-shot results for both CharacterGLM-3 and Emohaa. For the rest LLMs, we report both zero-shot and 4-shot results.

ChatEMO. We further conduct supervised fine-tuning of the GLM-4-9B-Chat model on both two tasks to develop our model. We conducted experiments on 8 Nividia RTX A6000 GPUs, setting the total batch size as 64, with gradient accumulation as 4, the maximum learning rate as 1e-5, and the total number of training epochs as 3.

4.3 Evaluation Metrics

Metric for Task1. According to previous studies (Chen et al., 2022a; Wen et al., 2021; Shen et al., 2020a), we used general metrics including Accuracy, Macro F1, Precision, and Recall to measure LLMs' performance on the emotion prediction task. Moreover, to address the limitation of general metrics that treat emotions discretely and overlook their nuanced spectrum-like nature, we introduced CAT-Dist as an additional metric to capture subtle similarities more effectively. CAT-Dist is inspired by the study by Smith and Ellsworth (Smith and Ellsworth, 1985), which mapped 15 emotions on the six dimensions (of the appraisal process). Each emotion is represented as a six-dimensional vector, and the specific values can be found in detail in Appendix A.6. We define CAT-Dist as the average Manhattan distance between two emotion vectors across all six dimensions:

$$\text{CAT-Dist}(E_1, E_2) = \frac{1}{D}\sum_{i=1}^{D}|E_{1,i} - E_{2,i}|$$

where $E_1 = (E_{1,1}, E_{1,2}, \ldots, E_{1,D})$ and $(E_2 = (E_{2,1}, E_{2,2}, \ldots, E_{2,D})$ are the two emotion vectors, D is the number of dimensions (here, $D = 6$).

Metric for Task2. Following previous studies (Varshney et al., 2021; Chen et al., 2022a), we choose the metrics below: (1) BLEU-1/2 (Papineni et al., 2002) that evaluates the model's output quality by comparing it to references through 1/2-gram matching; (2) ROUGE-1/2/L (Lin, 2004) that measures the longest common subsequence between output and reference, and (3) BERTScore (Zhang et al., 2019) which assesses output quality by cosine similarity of BERT embeddings.

Models	Setting	Acc. ↑	F1 ↑	Precision↑	Recall ↑	CAT-Dist ↓
GPT-4-0613	zero-shot	0.21	0.14	0.18	0.14	0.218
GLM-4-0520		0.18	0.12	0.24	0.13	0.229
DeepSeek V2		0.22	0.18	0.21	0.19	0.221
Qwen-2.5-72B		0.17	0.13	0.14	0.13	0.247
Llama3-8B-Chinese		0.12	0.07	0.08	0.11	0.265
GLM-4-9B		0.12	0.09	0.09	0.11	0.268
GPT-4-0613	4-shot prompting	0.25	0.16	0.20	0.16	0.213
GLM-4-0520		<u>0.27</u>	0.19	0.26	0.19	<u>0.207</u>
DeepSeek V2		0.26	<u>0.20</u>	**0.30**	<u>0.21</u>	0.213
Qwen-2.5-72B		0.24	0.19	0.20	0.20	0.224
Llama3-8B-Chinese		0.14	0.06	0.10	0.08	0.268
GLM-4-9B		0.17	0.10	0.11	0.11	0.242
ChatEMO	fine-tuning	**0.36**	**0.28**	<u>0.29</u>	**0.28**	**0.181**

Table 3: Performance comparison on Emotion Prediction. (**Bold** for best, <u>underline</u> for the second best).

Models	Setting	BLEU-1 ↑	BLEU-2 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BERTScore ↑
GPT-4-0613	zero-shot	23.63	5.13	22.71	2.66	19.10	61.07
GLM-4-0520		22.46	4.76	24.24	3.66	19.76	61.61
DeepSeek V2		24.60	5.38	24.49	3.45	20.11	61.36
Qwen-2.5-72B		24.45	5.36	23.48	3.46	19.46	61.55
Llama3-8B-Chinese		14.57	3.30	17.20	1.55	13.45	57.23
Emohaa		17.91	3.88	22.72	2.68	17.43	60.27
CharacterGLM3		25.67	**5.84**	23.89	3.53	19.94	61.43
GLM-4-9B		24.06	5.55	21.49	2.20	17.90	60.39
GPT-4-0613	4-shot prompting	**27.16**	<u>5.74</u>	25.52	3.68	21.39	62.40
GLM-4-0520		25.11	5.42	24.63	3.52	20.46	61.66
DeepSeek V2		25.83	5.47	**26.40**	4.22	<u>21.86</u>	62.26
Qwen-2.5-72B		26.02	5.53	25.92	<u>4.37</u>	21.78	<u>62.41</u>
Llama3-8B-Chinese		14.29	3.26	17.08	1.34	13.67	56.59
GLM-4-9B		23.43	5.44	22.19	2.20	18.11	60.78
ChatEMO	fine-tuning	<u>26.46</u>	5.59	<u>26.36</u>	**4.90**	**22.35**	**62.83**

Table 4: Performance comparison on generating emotional utterance. (**Bold** for best, <u>underline</u> for the second best).

5 Results

5.1 Emotion Prediction

The results of the emotion prediction task are shown in Table 3. Larger LLMs (i.e., GPT-4-0613, GLM-4-0520, DeepSeek V2, and Q-2.5-72B) perform better than smaller LLMs (i.e., Llama3-8B-Chinese and GLM-4-9B), due to more parameters and extensive pre-training data. Among large models, DeepSeek V2 achieves the highest accuracy and F1 score in zero-shot settings, while GLM-4-0520 shows the highest performance boost in few-shot settings, with a remarkable 50% increase in accuracy and 58.3% increase in F1 score. ChatEMO achieves the highest accuracy and F1 score, surpassing all baselines in zero-shot and few-shot settings. Supervised fine-tuning yields a twofold increase in both accuracy and F1 score compared to the original GLM-4-9B, demonstrating that our dataset provides comprehensive information necessary for precise emotion prediction.

5.2 Next Utterance Prediction

Results of this task are shown in Table 4. Similar to the emotion prediction task, large LLMs continue to outperform small ones in predicting the next utterance. However, CharacterGLM-3, a 6B model, shows comparable performance to large models in both zero-shot and few-shot settings. This may be attributed to its specialized design for human-like role-playing. Conversely, Emohaa's suboptimal performance may stem from its emphasis on providing empathetic support, highlighting differences between giving empathetic responses and expressing emotional utterances. ChatEMO performs the first or second-best across most metrics. Its proficiency lies in ROUGE-1/2/L metrics over BLEU-1/2, indicating precision in word identification but a tendency to generate irrelevant words. Notably, ChatEMO ranks among the top performers in BERTScore, suggesting a high degree of similarity in content between its outputs and the ground truth.

Models	SettingMatch (1-5)	Coherence (1-5)	Fluency (1-5)
GPT-4-0613	4.52	4.51	3.85
CharacterGLM-3	4.11	4.75	4.03
DeepSeek V2	3.16	3.48	3.52
ChatEMO	**4.55**	**4.80**	**4.52**

Table 5: Human ratings on generated utterances.

Given the subjective nature of utterance evaluation, the metrics above may not fully reflect the quality of generated responses. Therefore, we try to enhance our evaluation by randomly selecting 200 utterances from top-performing baselines and recruiting workers to evaluate them. ChatEMO, GPT-4-013, CharacterGLM-3, and DeepSeek V2 were chosen as they achieve the best performance on at least one metric. We incorporate the dimensions including SettingMatch, Coherence, and Fluency from the dialogue quality evaluation section to guide the human rating process. The results can be found in Table 5. To better demonstrate the comparison, we provide an example of generated utterances by the four LLMs in Appendix A.7.

6 Analysis

6.1 Ablation: How Our Framework Design Enhances Dialogue Quality?

We investigated how our data-generative framework enhances emotional dialogue quality. We omit key components of our framework to evaluate data quality without certain steps, comparing three versions: the full framework, one excluding belief and knowledge generation, and one without the appraisal process. Three Chinese-speaking workers are recruited to assess the quality of dialogues, based on five dimensions used in previous studies (Liu et al., 2024a; Zhang et al., 2024a; Ou et al., 2023; Zhong et al., 2022): (1) *coherence*, how much the dialogue maintains logical consistency without confusion; (2) *naturalness*, how well the dialogue aligns with Chinese speech habits; (3) *correctness*, whether the label belongs to the given 15 emotions; (4) *contextual relevance*, the degree to which the dialogue aligns with the settings; and (5) *emotional relevance*, the alignment of emotional utterances with the settings. A rating guideline is provided for more reliable ratings, which is listed in the Appendix A.8.

Human evaluation result (Table 6) demonstrates that the absence of belief and knowledge highly impacts coherence and naturalness, as they contribute useful information to enrich the storyline and elevate the generated dialogues' quality. Moreover, omitting the appraisal process leads to notable decreases in correctness, contextual, and emotional relevance. This is because the appraisal process offers detailed guidelines for generating appropriate emotions and actions, ensuring that emotions align logically and dialogues fit the context.

Dimensions	Score Range	CAT-BEAR	w/o Belief & Knowledge	w/o Appraisal
Coherence	1-5	4.81	4.29 (↓11%)	4.32 (↓10%)
Naturalness	1-5	4.66	4.02 (↓14%)	4.14 (↓11%)
Correctness	0-1	1.00	0.66 (↓34%)	0.50 (↓50%)
Contextual Relevance	1-5	4.90	4.28 (↓13%)	4.18 (↓15%)
Emotional Relevance	1-5	4.77	3.97 (↓17%)	3.89 (↓18%)

Table 6: Human ratings on generated dialogues by different data synthetic frameworks.

6.2 How Emotion Labels Impact the Quality of Generated Utterances?

To explore whether an agent can first self-predict emotion before generating a response—thereby enhancing emotional response generation—we draw on the emotion chain-of-thought method (Li et al., 2024). Our experiments consist of two parts: (1) *Conditional Setting*, where we examine whether providing the ground-truth emotion label improves the quality of the generated utterance; and (2) *Joint-Modeling Setting*, which encourages the model to first predict the emotion and then generate the response. We perform supervised fine-tuning for both settings to align as above. As shown in Table 7, when the emotion label is accurate, the quality of generated utterances significantly improves. Additionally, joint modeling not only enhances emotion labeling accuracy but also results in contextually relevant and emotionally resonant responses. This implies that future work could leverage our dataset to develop a more sophisticated emotion CoT, further enhancing agent emotional intelligence.

Models	Acc. ↑	F1 ↑	CAT-Dist ↓
ChatEMO	0.36	0.28	0.181
joint	0.39(↑8%)	0.38(↑36%)	0.173(↓4%)
Models	**BLEU-1/2 ↑**	**ROUGE-1/2/L ↑**	**BERTScore ↑**
ChatEMO	26.46 / 5.59	26.36 / 4.90 / 22.35	62.83
cond	**29.82 / 6.09**	**29.49 / 6.75 / 25.02**	**64.58**
joint	28.28 / 5.93	28.03 / 5.75 / 23.76	63.61

Table 7: Effects of emotion label prediction accuracy on response quality. (task1 above, task2 below).

7 Conclusions

In this paper, we introduced the Emotion Conversations Generation Framework, which harnesses the Cognitive Appraisal Theory to generate emotionally appropriate responses in conversational agents. Our framework delves into intra-individual factors such as situational construal and personality, incorporating a comprehensive appraisal process to enable nuanced emotion predictions and utterances generation. Building upon this framework, we curated a multi-turn conversational corpus, CAPE, comprising 2,848 emotional dialogues meticulously refined by human crowd workers. Utilizing CAPE, we fine-tuned a model and conducted experiments that showcased the superior performance of our fine-tuned LLM over various robust baseline models in both emotion prediction and utterance generation tasks. The results of our study underscore the transformative potential of integrating cognitive appraisal insights into LLMs, highlighting the substantial enhancements in accuracy and effectiveness attainable in automated emotional interactions.

Limitations

This study employed GPT-4-turbo-0409 to generate a Chinese conversational dataset. The language proficiency of GPT-4-turbo in Chinese may affect the fluency and naturalness of the agents' utterances. Despite instructing GPT-4-turbo to create conversations that align with Chinese style and culture, some issues may still arise due to cultural differences. We also recognize the value of a larger-scale dataset but were constrained by the high costs of GPT-4-turbo data generation and human quality control, limiting our dataset size. Future work could address this limitation by exploring more cost-effective data generation and validation methods. Moreover, due to computational costs, this study only fine-tuned one LLM, GLM-4-9B, to evaluate the effectiveness of CAPE. We hope that future studies will apply more extensive training to CAT-BEAR and CAPE to address this issue.

Ethics Statement

In conducting this research on emotion generation in agent conversations, we prioritize ethical considerations to ensure the responsible development and deployment of our CAT-BEAR framework. Our commitment involves adhering to principles of transparency, privacy, and fairness throughout the study. We ensure that the multi-turn conversational corpus, CAPE, utilized in training and evaluation, aligns with ethical standards for data collection and does not contain personal or sensitive information. Furthermore, we conduct both automatic and manual evaluations with due consideration of potential biases and the diverse representation of emotions. When hiring workers to review and adjust the dataset, we ensure that dialogues generated by GPT-4-turbo are free from personal privacy concerns, political bias, and similar issues. The introduction of the CAT-Dist metric aims to provide an objective evaluation of emotional appropriateness without infringing on subjective human experiences.

Acknowledgements

The authors would like to express their gratitude to AI iPal (Aixiaoban) for their support and valuable contributions to this research.

References

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Scott Brave and Cliff Nass. 2007. Emotion in human-computer interaction. In *The human-computer interaction handbook*, pages 103–118. CRC Press.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022a. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022b. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2022. Pal: Persona-augmented emotional support conversation generation. *arXiv preprint arXiv:2212.09235*.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *arXiv preprint arXiv:2406.11161*.

Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. 2024. An appraisal-based chain-of-emotion architecture for affective language model game agents. *Plos one*, 19(5):e0301033.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Jiawen Deng and Fuji Ren. 2021. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67.

Elaine S Elliott and Carol S Dweck. 1988. Goals: an approach to motivation and achievement. *Journal of personality and social psychology*, 54(1):5.

Yumeng Fu. 2024. Ckerc: Joint large language models with commonsense knowledge for emotion recognition in conversation. *arXiv preprint arXiv:2403.07260*.

David C Funder. 2016. Taking situations seriously: The situation construal model and the riverside situational q-sort. *Current Directions in Psychological Science*, 25(3):203–208.

Ruyi Gan, Ziwei Wu, Renliang Sun, Junyu Lu, Xiaojun Wu, Dixiang Zhang, Kunhao Pan, Ping Yang, Qi Yang, Jiaxing Zhang, et al. 2023. Ziya2: Data-centric learning is all llms need. *arXiv preprint arXiv:2311.03301*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Daniel Goleman. 2020. *Emotional intelligence: Why it can matter more than IQ*. Bloomsbury Publishing.

Katharine H Greenaway, Elise K Kalokerinos, and Lisa A Williams. 2018. Context is everything (in emotion research). *Social and Personality Psychology Compass*, 12(6):e12393.

James J Gross and Lisa Feldman Barrett. 2011. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion review*, 3(1):8–16.

Craig A Hill. 1987. Affiliation motivation: people who need people... but in different ways. *Journal of personality and social psychology*, 52(5):1008.

Ruud Hortensius, Felix Hekele, and Emily S Cross. 2018. The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):852–864.

Hangyeol Kang, Maher Ben Moussa, and Nadia Magnenat-Thalmann. 2024. Nadine: An llm-driven intelligent social robot with affective capabilities and human-like memory. *arXiv preprint arXiv:2405.20189*.

Richard S Lazarus. 1991. *Emotion and adaptation*. Oxford University Press.

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. 2022. C3kg: A chinese commonsense conversation knowledge graph. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1369–1383.

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kristen A Lindquist. 2017. The role of language in emotion: existing evidence and future directions. *Current opinion in psychology*, 17:135–139.

Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. 2024a. Speak from heart: An emotion-guided llm-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 533–542.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. 2024b. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *arXiv preprint arXiv:2401.08508*.

John C Masters. 1991. Strategies and mechanisms for the personal and social control of emotion. *The development of emotion regulation and dysregulation*, 338:182–207.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, Zhongyuan Wang, and Kun Gai. 2023. Dialogbench: Evaluating llms as human-like dialogue systems. *arXiv preprint arXiv:2311.01677*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W Schuller. 2024. Customising general large language models for specialised emotion recognition tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11326–11330. IEEE.

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023. Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv preprint arXiv:2301.04907*.

Jay Ratican and James Hutson. 2023. The six emotional dimension (6de) model: A multidimensional approach to analyzing human emotions and unlocking the potential of emotionally intelligent artificial intelligence (ai) via large language models (llm). *Journal of Artificial Intelligence and Robotics*, 1(1).

Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. 2023. A chatbot for mental health support:

exploring the impact of emohaa on reducing mental distress in china. *Frontiers in digital health*, 5:1133987.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Rüdiger J Seitz and Hans-Ferdinand Angel. 2020. Belief formation–a driving force for brain evolution. *Brain and Cognition*, 140:105548.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020a. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on multimedia*, pages 493–502.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020b. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on multimedia*, pages 493–502.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Renliang Sun, Mengyuan Liu, Shiping Yang, Rui Wang, Junqing He, and Jiaxing Zhang. 2024. Fostering natural conversation in large language models with nico: a natural interactive conversation dataset. *arXiv preprint arXiv:2408.09330*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Deeksha Varshney, Asif Ekbal, and Pushpak Bhattacharyya. 2021. Modelling context emotions using multi-task learning for emotion controlled dialog generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2919–2931.

Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. 2024. Llama3-8b-chinese-chat (revision 6622a23).

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020.

T Mikael Winberg, Jenny M Hellgren, and Torulf Palm. 2014. Stimulating positive emotional experiences in mathematics learning: influence of situational and personal factors. *European Journal of Psychology of Education*, 29:673–691.

Frank Xing. 2024. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024. Enhancing empathetic response generation by augmenting llms with small-scale empathetic models. *arXiv preprint arXiv:2402.11801*.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024a. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19515–19524.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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*.

Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W Schuller. 2024b. Refashioning emotion recognition modelling: The advent of generalised large models. *IEEE Transactions on Computational Social Systems*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint arXiv:2205.10237*.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023a. Charactergl m: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023b. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. *arXiv preprint arXiv:2307.07994*.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2022. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. *arXiv preprint arXiv:2208.08845*.

A Appendix

A.1 Situational construals' list

Here we present 89 situational construals, as illustrated in Table 8, encompassing various topics related to three fundamental elements: individuals involved, the specific situation, and corresponding behaviors. Each situational construal is used to generate 32 dialogues, incorporating diverse personal factors to enhance the dialogue's storyline. To enrich the narrative of the dialogues, we elaborate on the situational construal for each dialogue based on individual characteristics. For instance, as depicted in Figure 2, the situational construal "talking is permitted" is expanded to "AA and BB are taking a walk in the park, discussing the latest popular movies, sharing their afterthoughts."

A.2 Guideline for GPT-4 to generate belief and knowledge

We provide the prompt depicted in Figure 4 and the intra-individual factors of the two speakers to GPT-4-turbo-0409, to generate each speaker's beliefs and knowledge. The generated beliefs encompass empirical, relational, and conceptual beliefs, along with knowledge.

A.3 Guideline for GPT-4 to find the most appropriate emotion

Figure 5 depicts the guideline offered to GPT-4-turbo-0409 for emotion prediction through the appraisal process, as outlined in stage two of Figure 2. It elucidates the definition of each dimension of the appraisal process, such as unpleasantness, effort, attention, certainty, control, and responsibility, and teaches GPT-4-turbo to sequentially categorize the degree of each dimension to determine the most likely emotion label and produce an emotion-aligned utterance.

A.4 Guideline for human evaluation of data quality

In data quality control, six Chinese-speaking workers are provided with the guideline in Figure 6 to assess the quality of both raw and human-refined dialogues at the utterance level. Each rater was presented with an utterance, its corresponding emotion label, the dialogue history preceding the utterance, and the intra-individual factors of the two speakers for comprehensive evaluation.

A.5 Representative Emotional Dialogue Datasets Comparison

Table 9 presents a comparison between CAPE and other existing emotion dialogue datasets, focusing on language, the range of emotion categories covered, as well as the number of situational and personal factors considered. Notably, CAPE boasts the most extensive coverage across emotions, situational contexts, and personal attributes. Furthermore, in contrast to characters in TV series, which often have predefined characters, the synthetic characters in our dataset offer a more diverse range of backgrounds, paving the way for enhanced data generation capabilities.

A.6 Emotions' scores on six dimensions

Table 10 presents the original scores reflecting the performance of each emotion across six appraisal dimensions, as outlined in the research conducted by Smith and Ellsworth (Smith and Ellsworth, 1985). This data is normalized by dimension, with each emotion represented as a normalized six-dimensional vector. The CAT-Dist, defined in the evaluation metrics section, is calculated as the Manhattan distance between two emotion vectors.

A.7 Comparison of generated utterances by top-performing LLMs

As shown in Figure 1, we select one representative example of the generated responses by top-performing models in the second task. The figure includes intra-individual factors and dialogue history for context. The examples illustrated in the figure demonstrate that the utterance generated by ChatEMO closely aligns with the ground truth in terms of content, reflecting BB's solitary and critical personality, alongside the belief that individual performance is more important than teamwork.

中文	**English**
1. 该情境有令人愉快的可能性	1. Situation is potentially enjoyable.
2. 该情境情况复杂	2. Situation is complex.
3. 该情境中有某项工作需要完成	3. A job needs to be done.
4. 该情境中有人尽力使你对他（或她）留下深刻印象	4. Someone is trying to impress P.
5. 该情境中有人试图让你相信某事	5. Someone is trying to convince P of something.
6. 该情境中需要依靠你去做某事	6. P is counted on to do something.
7. 该情境中允许讲话	7. Talking is permitted.
8. 该情境期望甚至要求人们说话	8. Talking is expected or demanded.
9. 该情境中你被要求做某事	9. P is being asked for something.
10. 该情境中有人需要帮助	10. Someone needs help.
11. 该情境中小细节很重要	11. Minor details are important.
12. 该情境涉及到生活方式或政治信仰等方面的价值观	12. Situation evokes values concerning lifestyles or politics.
13. 该情境提供了一个展示才智的机会	13. Affords an opportunity to demonstrate intellectual capacity.
14. 该情境具有不确定性	14. Situation is uncertain.
15. 该情境中在场的或是被谈及的另一个人正受到威胁	15. Another person (present or discussed) is under threat.
16. 该情境中你遭到直接或间接地批评	16. P is being criticized, directly or indirectly.
17. 该情境中有人试图要支配或领导你	17. Someone is attempting to dominate or boss P.
18. 该情境十分有趣好玩	18. Situation is playful.
19. 该情境有可能引起自我反省	19. Introspection is possible.
20. 该情境中事情发生得很快	20. Things are happening quickly.
21. 该情境中在场的或是被谈及的某一个人心情不好甚至是非常痛苦	21. Someone (present or discussed) is unhappy or suffering.
22. 该情境中有另一位可靠的人在场	22. A reassuring other person is present.
23. 该情境中你因为某事受到指责	23. P is being blamed for something.
24. 该情境中需要做出决策	24. A decision needs to be made.
25. 该情境中要求理性思考	25. Rational thinking is called for.
26. 该情境中要求自制力或对自我的约束	26. Situation calls for self-restraint.
27. 该情境包含竞争	27. Situation involves competition.
28. 该情境提供给你一个机会做讨人喜欢的事情	28. Affords an opportunity for P to do things that might make P liked or accepted.
29. 该情境中有人在寻求认同或支持	29. Others are present who need or desire reassurance.
30. 该情境中有令人沮丧的情况出现	30. Situation entails frustration.
31. 该情境与你的外表吸引力有关	31. Physical attractiveness of P is relevant.
32. 该情境中给人留下好印象对你很重要	32. It is important for P to make a good impression.
33. 该情境会令一些人感到紧张和不安	33. Situation would make some people tense and upset.
34. 该情境涉及一个或多个小麻烦	34. Situation includes one or more small annoyances.
35. 该情境可能唤起温情或同情心	35. Situation might evoke warmth or compassion.
36. 该情境中某个人或某项活动可能遭到诋毁或暗中破坏	36. A person or activity could be undermined or sabotaged.
37. 该情境中你有可能欺骗某人	37. It is possible for P to deceive someone.
38. 该情境中除你之外的其他人可能有欺诈之意	38. Someone else in this situation (other than P) might be deceitful.
39. 该情境可能引发敌对情绪	39. Situation may cause feelings of hostility.
40. 该情境中人们对某件事的看法产生分歧	40. People are disagreeing about something.
41. 该情境提供了一个发表独特见解或思想的机会	41. Affords an opportunity to express unusual ideas or points of view.
42. 该情境包含人身威胁	42. Situation contains physical threats.
43. 该情境包含对情绪或情感的威胁	43. Situation contains emotional threats.
44. 该情境引发道德或伦理问题	44. Situation raises moral or ethical issues.
45. 该情境要求快速决策或迅速行动	45. A quick decision or quick action is called for.
46. 该情境中可以表达任何一种情感	46. Situation allows a free range of emotional expression.
47. 该情境中其他当事者可能有相冲突的或是刻意隐藏的目的或动机	47. Others present might have conflicting or hidden motives.
48. 该情境导致或可能导致压力或创伤	48. Situation entails or could entail stress or trauma.
49. 该情境提供了一个沉思、空想，或者幻想的机会	49. Affords an opportunity to ruminate, daydream or fantasize.
50. 该情境有引发你负疚感的可能性	50. Situation has potential to arouse guilt in P.
51. 该情境中出现或可能发展出亲密的个人关系	51. Close personal relationships are present or have the potential to develop.
52. 该情境中要依靠除你以外的某一个人去做某一件事	52. Someone other than P is counted on to do something.
53. 该情境包含对智力或认知的刺激	53. Situation includes intellectual or cognitive stimuli.
54. 该情境中实现目标需要相当坚定的自信	54. Assertiveness is required to accomplish a goal.
55. 该情境包含对某种欲望即时满足的可能性	55. Situation includes potential for immediate gratification of desires.
56. 该情境中可能出现人际互动	56. Social interaction is possible.
57. 该情境是诙谐幽默的或是有幽默因素的	57. Situation is humorous or potentially humorous.
58. 该情境中你是引人注目的焦点	58. P is the focus of attention.
59. 该情境包含感官刺激	59. Situation includes sensuous stimuli.
60. 该情境关乎你的身体健康	60. Situation is relevant to bodily health of P.
61. 该情境中成功需要对自我的深入剖析	61. Success in this situation requires self-insight.
62. 该情境中你控制着他人所需要的资源	62. P controls resources needed by others.
63. 该情境中的别人展现出了大量与人际关系有关的线索	63. Others present a wide range of interpersonal cues.
64. 该情境包括对行为的限制	64. Situation includes behavioral limits.
65. 该情境包含审美刺激	65. Situation includes aesthetic stimuli.
66. 该情境有增加焦虑感的可能性	66. Situation is potentially anxiety-inducing.
67. 该情境包含对你的直接或间接的要求	67. Situation makes demands on P.
68. 该情境提供了一个表达或证明抱负的机会	68. Affords an opportunity to express or demonstrate ambition.
69. 该情境可能会让你感到自己能力不够	69. Situation might make P feel inadequate.
70. 该情境包含一些可从性爱的角度诠释的刺激	70. Situation includes stimuli that could be construed sexually.
71. 该情境中形势要求变化很快	71. Situational demands are rapidly shifting.
72. 该情境中你被辱骂或是被伤害	72. P is being abused or victimized.
73. 该情境中出现异性成员	73. Members of the opposite sex are present.
74. 该情境中出现了可能成为你恋人的对象	74. Potential romantic partners for P are present.
75. 该情境可能会唤起内心的冲突	75. Situation has potential to arouse competing motivations.
76. 该情境基本上简单明了	76. Situation is basically simple and clear-cut.
77. 该情境提供了一个展现个人魅力的机会	77. Affords an opportunity to express charm.
78. 该情境涉及人跟人之间的比较	78. Situation involves social comparison.
79. 该情境涉及到权力的问题	79. Situation raises issues of power.
80. 该情境提供了一个展现男性阳刚一面的机会	80. Affords an opportunity to express masculinity.
81. 该情境中他人可能需要你的建议或向你征求意见	81. Others may need or are requesting advice from P.
82. 该情境中你的独立性或自主权受到质疑或威胁	82. Independence or autonomy of P is questioned or threatened.
83. 该情境可能激发某些特定的情绪或情感	83. Situation is potentially emotionally arousing.
84. 该情境提供了一个证明口才的机会	84. Affords an opportunity for demonstrating verbal fluency.
85. 该情境中当事者的社会角色或地位等级各不相同	85. People who are present occupy different social roles or levels of status.
86. 该情境中你被迫随大流	86. P is being pressured to conform to the actions of others.
87. 该情境中成功需要合作	87. Success requires cooperation.
88. 该情境中你受到恭维或称赞	88. P is being complimented or praised.
89. 该情境提供了一个展现女性阴柔一面的机会	89. Affords an opportunity to express femininity.

Table 8: Situational Construals' List in Chinese and English

System Prompt:

给定情境，补充两个角色的信念和知识，这些需要和情境以及两个人的个人情况有关系。
Given the situation, generate the beliefs and knowledge of the two speakers, which need to be related to the situation and the two persons' intra-individual factors.

信念指的是，在特定情况下，什么是规范上适当的、可行的、合法的、可原谅的。三种信念和情境以及两个人的个人情况都有非常大的关系。
有三种不同的信念类型：
经验性信念（物体、感知价值、感知某物、相信什么）
关系性信念（事件，相信某个人，和归属目标倾向有关系）
概念性信念（叙述性，相信）
There are three different types of beliefs:
Empirical belief (to believe objects or perceived value)
Relational belief (to believe someone, related to affiliation goals)
Conceptual beliefs (to believe in narratives)

知识指的是，与目标倾向相关的东西，以及什么是好的有益的，什么是不好的有危害的。知识和情境以及两个人的个人情况都有非常大的关系。
Knowledge is defined as the information that is related to goals and encompasses the understanding of what may cause harm or provide benefits. Knowledge is highly related to the situation and the two persons' intra-individual factors.

Figure 4: Guideline for GPT-4 to generate belief and knowledge

Datasets	Source	Lang.	Emotion Coverage	Situations	Personal Factors
EmoryNLP (Zahiri and Choi, 2018)	TV-series	EN	7 (sad, mad, scared, powerful, peaceful, joyful, and neutral)	Not clarified	Not clarified
MEmoR (Shen et al., 2020b)	TV-series	EN	14 (joy, anger, disgust, sadness, surprise, fear, anticipation...)	Not clarified	16PF, MBTI, Big FIVE
CPED (Chen et al., 2022b)	TV-series	CN	13 (happy, grateful, relaxed, other-positive neutral, angry...)	10 scenes	Gender, Age, Big FIVE
M3ED (Zhao et al., 2022)	TV-series	CN	7 (happy, surprise, sad, disgust, anger, fear, and neutral)	Not clarified	Gender, Age
CAPE (Ours)	Synthetic	CN	15 (happy, surprise, sadness, pride, challenge...)	89 seed situations	Big FIVE, Goals, Belief, Knowledge

Table 9: Comparison of CAPE with other representative emotion dialogue datasets.

Emotion	*Unpleasantness*	*Effort*	*Attention*	*Certainty*	*Control*	*Responsibility*
Happiness	-1.46	-0.33	0.15	-0.46	-0.21	0.09
Sadness	0.87	-0.14	-0.21	0.0	1.15	-0.36
Anger	0.85	0.53	0.12	-0.29	-0.96	-0.94
Boredom	0.34	-1.19	-1.27	-0.35	0.12	-0.19
Challenge	-0.37	1.19	0.52	-0.01	-0.2	0.44
Hope	-0.50	-0.18	0.31	0.46	0.35	0.15
Fear	0.44	0.63	0.03	0.73	0.59	-0.17
Interest	-1.05	-0.07	0.70	-0.07	0.41	-0.13
Contempt	0.89	-0.07	0.80	-0.12	-0.63	-0.50
Disgust	0.38	0.06	-0.96	-0.39	-0.19	-0.50
Frustration	0.88	0.48	0.60	-0.08	0.22	-0.37
Surprise	-1.35	-0.66	0.40	0.73	0.15	-0.94
Pride	-1.25	-0.31	0.02	-0.32	-0.46	0.81
Shame	0.73	0.07	-0.11	0.21	-0.07	1.31
Guilt	0.60	0.0	-0.36	-0.15	-0.29	1.31

Table 10: Emotion's Scores on Six Dimensions.

A.8 Guideline for human evaluation of the quality of dialogues generated by different frameworks

Figure 8 illustrates the guideline that we provided to an additional three Chinese-speaking workers for rating the generated dialogues produced by three versions of frameworks. In contrast to the data quality control phase which rates data by utterance, this guideline is at the dialogue level. Raters will receive intra-individual factors and the complete dialogue context when conducting their assessment.

A.9 Extension of ERC task

To show the value of Cognitive Appraisal Theory, we further include emotion recognition in conversations (ERC) tasks on the CPED bench-

System Prompt:

以下是情绪判断方法，你将基于6个维度来预测在这个情境下该角色的情绪。每个维度的定义如下：
Below is the method to predict emotions, where you will predict the emotion of the user based on 6 dimensions. Each dimension is defined below:

不愉快的程度：越高/正的分数表示越不愉快。
Unpleasantness: a higher/positive score indicates higher unpleasantness.
预期可能付出的努力：高分表明预期中付出的努力会很多。
Effort, a high score indicates that a lot of effort is expected.
注意力：高分表明注意力活动增加。
Attention: a high score indicates increased attentional activity.
不确定性：得分高表明不确定性增加。
Certainty: a high score indicates increased uncertainty.
不可控性：事物可以被说话者控制/被人类控制，则得分低，如果事物不能被说话者控制，而是被情景控制，则得分高，得分高表明事情被外界控制多。
Control: if things can be controlled by the speaker/by humans, the score is low; otherwise, the score is high, indicating that things are controlled more by the outside world.
责任：如果引起情绪的事件是由其他人导致的，则得分低，如果是事件由自己负责，则得分高，得分高表明自我的责任强。
Responsibility: low score if the event causing the emotion is caused by someone else, high score if the event is by the responsibility of the self, high score indicates more responsibility by the self.

(1) 先根据不愉快的程度分类：
如果不愉快的程度高，那么你可以选择的情绪范围是：悲伤，愤怒，无聊，恐惧，蔑视，厌恶，挫折，羞耻，内疚
如果不愉快的程度低，那么你可以选择的情绪范围是：快乐，挑战，希望，兴趣，惊讶，骄傲
(1) Start by categorizing according to the degree of unpleasantness:
If the degree of unpleasantness is high, then the range of emotions you can choose from are: sadness, anger, boredom, fear, contempt, disgust, frustration, shame, guilt
If the degree of unpleasantness is low, then the range of emotions you can choose from is: joy, challenge, hope, interest, surprise, pride

(2) 再根据预期可能付出的努力分类：
如果预期中付出的努力会多，那么你可以选择的情绪范围是：愤怒，挑战，恐惧，厌恶，挫折，羞耻，内疚
如果预期中付出的努力会不多，那么你可以选择的情绪范围是：快乐，悲伤，无聊，希望，兴趣，蔑视，惊讶，骄傲
(2) Then categorize them according to the expected effort:
If the expected effort will be high, then you can choose from the following range of emotions: anger, challenge, fear, disgust, frustration, shame, guilt.
If it is expected that there will be little effort, then the range of emotions you can choose from is: happiness, sadness, boredom, hope, interest, contempt, surprise, pride.

(3) 再根据注意力分类：
如果可能的注意力活动增加，那么你可以选择的情绪范围是：快乐，愤怒，挑战，希望，恐惧，兴趣，蔑视，挫折，惊讶，骄傲
如果可能的注意力活动减少，那么你可以选择的情绪范围是：悲伤，无聊，厌恶，羞耻，内疚
(3) Then categorize according to attention:
If there is an increase in possible attentional activity, then the range of emotions you can choose from are: happy, angry, challenged, hopeful, fearful, interested, defiant, frustrated, surprised, proud
If there is a decrease in possible attentional activity then the range of emotions you can choose from are: sadness, boredom, disgust, shame, guilt

(4) 再根据不确定性分类：
如果不确定性增加，那么你可以选择的情绪范围是：悲伤，希望，恐惧，惊讶，羞耻
如果不确定性减少，那么你可以选择的情绪范围是：快乐，愤怒，无聊，挑战，兴趣，蔑视，厌恶，挫折，骄傲，内疚
(4) Then categorize based on uncertainty:
If uncertainty increases, then the range of emotions you can choose from are: sadness, hope, fear, surprise, shame
If uncertainty decreases, then the range of emotions you can choose from are: happiness, anger, boredom, challenge, interest, contempt, disgust, frustration, pride, and guilt.

(5) 再根据不可控性分类：
如果事物被外界控制多，那么你可以选择的情绪范围是：悲伤，无聊，希望，恐惧，兴趣，挫折，惊讶
如果事物被自己控制多，那么你可以选择的情绪范围是：快乐，愤怒，挑战，蔑视，厌恶，羞耻，骄傲，内疚
(5) Then categorize according to control:
If things are controlled more by the outside world, then the range of emotions you can choose from are: sadness, boredom, hope, fear, interest, frustration, surprise
If things are controlled by yourself, then you can choose from a range of emotions: happiness, anger, challenge, contempt, disgust, shame, pride, guilt.

(6) 再根据责任分类：
如果自我的责任强，那么你可以选择的情绪范围是：快乐，挑战，希望，骄傲，羞耻，内疚
如果自我的责任弱，那么你可以选择的情绪范围是：悲伤，愤怒，无聊，恐惧，兴趣，蔑视，厌恶，挫折，惊讶
(6) Finally, categorize according to responsibility:
If the user has a strong degree of self-responsibility, then the range of emotions you can choose from are: happiness, challenge, hope, pride, shame, guilt
If the degree of self-responsibility is weak, then the range of emotions you can choose from is: sadness, anger, boredom, fear, interest, contempt, disgust, frustration, surprise.

如果到了任意一个维度，你发现你只剩下了一个选项，那么你可以选择这个情绪选项。
如果到了最后一步，你发现你还有多个选项，那么你可以随机选择其中一个选项。
If by the time you get to any of the dimensions, you realize that you are left with only one option, then you can choose that emotional option.
If by the last step, you find that you have more than one option left, then you can choose one of the options at random.

你只能在这15种情绪当中选择：快乐，悲伤，愤怒，无聊，挑战，希望，恐惧，兴趣，蔑视，厌恶，挫折，惊讶，骄傲，羞耻，内疚
You can only choose between these 15 emotions: [Happiness, Sadness, Anger, Boredom, Challenge, Hope, Fear, Interest, Contempt, Disgust, Frustration, Surprise, Pride, Shame, Guilt]

Figure 5: Guideline for GPT-4 to find the most appropriate emotion

mark (Chen et al., 2022b). We compared with two strong baselines: DialogXL (Shen et al., 2021) and BERT+AVG+MLP (Chen et al., 2022b). The results show that our trained model exhibits superior performance in the ERC tasks, demonstrating its ability to better capture the nuances and complexities of emotional expressions in dialogues.

Model	Neg. Acc	Neu. Acc	Pos. Acc	Avg. Acc	Macro F1
DialogXL	60.45	41.45	41.74	51.24	46.96
BERT+AVG+MLP	61.40	40.10	42.95	51.50	48.02
GLM-4-9B (CAPE)	**71.05**	**44.42**	**46.82**	**57.56**	**54.04**

Table 11: Emotion recognition in conversation (ERC) tasks on the CPED benchmark (Chen et al., 2022b).

话语标注手册

标注总目标 1. 认真阅读场景和人物设定 2. 在对应的设定下，为情绪的类型合理性和强度+语言内容的合理性和流畅性打分
数据由两个部分构成：人物场景设定对话历史 & 情绪标签 & 话语
打分标准
1. 情绪标签类型合理性（EmoCategory）：说话者的情绪标签类型是否符合人物场景设定
   0分=情绪类型不合适，1分=情绪类型合适
   情绪分类限定在15种：
   快乐，悲伤，愤怒，无聊，挑战，希望，恐惧，兴趣，蔑视，厌恶，挫折，惊讶，骄傲，羞耻，内疚
   相同情况下，可能有多种合理的情绪选择，只要我们的情绪标签属于合理的范围即可（即，不一定要选择最优的情绪，只要在合理的范围就行）
   案例：某人遇到了讨厌的事情，那么“厌恶”，“挫折”，“愤怒”等等都是合理的情绪，情绪标签属于其中一种，则计分=1
2. 内容与情绪标签类型一致（EmoMatch）：说话者言语内容是否符合情绪标签类型
   打分范围1-5分，1分=非常不符合（说话内容与情绪标签类型完全相反），3分=有点符合（存在些许与情绪标签不一致的内容），5分=非常符合
3. 内容与设定一致（SettingMatch）：除了情绪外，说话者言语内容是否符合人物场景设定
   打分范围1-5分，1分=非常不符合（说话内容违反了几乎所有的设定点），3分=有点符合（存在部分设定和说话内容不一致），5分=非常符合
4. 对应情绪标签类型强度（EmoIntensity）：说话者的话语内容中，对应情绪标签的情绪强度是多少（e.g.，如标签是<愤怒>，话语内容中愤怒的强度是多少）
   打分范围0-2分，0分=没有对应情绪，1分=少许对应情绪，2分=足够对应情绪
5. 语言合理性（Coherence）：说话者的话语内容是否适合当前对话的上下文逻辑（i.e.，不存在逻辑混乱，前言不搭后语）
   打分范围1-5分，1分=非常不合适（几乎都是逻辑混乱&前言不搭后语的情况），3分=有点合适（存在少量逻辑混乱），5分=非常合适
6. 语言流畅性（Fluency）：说话者的话语内容是否流畅和容易理解（i.e.，语法和口语化不存在太多问题）
   打分范围1-5分，1分=一点也不，3分=有点，5分=非常

Utterance Rating Manual

General objectives: 1. read the settings carefully 2. rate the reasonableness and intensity of the type of emotion + the reasonableness and fluency of the verbal content in the corresponding settings
The data consists of two parts: Settings & dialog history & emotion labels & utterances.
Scoring Criteria
1. EmoCategory: Whether the speaker's emotion label is appropriate for the character's scenario.
   0 = label inappropriate, 1 = lable appropriate.
   The label is limited to the following 15 categories:
   Happy, Sad, Angry, Bored, Challenged, Hopeful, Fearful, Interested, Scornful, Disgusted, Frustrated, Surprised, Proud, Shameful, Guilty.
   There may be more than one candidates for the same situation, as long as the emotion labels being reasonable (i.e., it is not required that the label must be the best one, as long as it looks reasonable)
   Example: if someone encounters something he hates, then “disgust”, “frustration”, “anger”, etc. are all reasonable emotions. If the emotion label belongs to one of them, then the score = 1.
2. EmoMatch: Whether the content of the speaker's utterance matches the type of emotion label.
   Scores range from 1-5, with 1 = very inconsistent (the content of the speaker's speech is the exact opposite of the emotion label type), 3 = somewhat consistent (there is some inconsistency with the emotion label), and 5 = very consistent.
3. SettingMatch: Whether the content of the speaker's utterance matches the settings.
   Score range 1-5, 1=very inconsistent (the content of the speech violates almost all of the setting points), 3=somewhat consistent (there are some inconsistencies between the setting and the content of the speech), 5=very consistent
4. EmoIntensity: the intensity of the emotion label in the speaker's content (e.g., if the label is <Anger>, what is the intensity of anger in the content).
   The scale is 0-2, with 0 = no emotion, 1 = little emotion, and 2 = enough emotion.
5. Coherence: Whether the content of the speaker's utterance fits into the logic of the current conversation (i.e., there is no logical confusion, no inconsistency).
   Score range 1-5, 1=very inappropriate (almost always confusing & inconsistent), 3=somewhat appropriate (a little confusing), 5=very appropriate.
6. Fluency: Whether the content of the speaker's words is fluent and easy to understand (i.e., grammar and colloquialisms are not too much of a problem).
   Score range 1-5, 1=not at all, 3=somewhat, 5=very much

Figure 6: Guideline for human evaluation of generated data quality.

Intra-individual factors:

AA
性格： 谨慎的，粗心的，外向的，友好的，敏感的
成就目标倾向： 低成就目标
归属目标倾向： 高隶属关系目标
经验性信念： 清理能提升公园环境美观
关系性信念： 团队合作能增进社区联系
概念性信念： 社区服务是个人责任和荣誉
知识： 社交活动对心理健康有益

BB
性格： 有创造力的，高效的，孤独的，挑剔的，敏感的
成就目标倾向： 高成就目标
归属目标倾向： 低隶属关系目标
经验性信念： 计划需具备高效执行可能
关系性信念： 个人表现优于团队合作
概念性信念： 成效是评价活动成功的关键
知识： 有效管理是成功的前提条件

AA
Personality: cautious, careless, outgoing, friendly, sensitive
Achievement goal: Low orientation towards
Affiliation goal: High orientation towards affiliation
Empirical belief: Cleaning up enhances the environment of the park
Relational belief: Teamwork enhances community connections
Conceptual belief: Community service is a responsibility and honor
Knowledge: Socialization is good for mental health

BB
Personality: inventive, efficient, solitary, critical, sensitive
Achievement goal: High orientation towards
Affiliation goal: Low orientation towards affiliation
Empirical belief: Plans need to have the potential for efficient execution.
Relational belief: Individual performance is more important than teamwork
Conceptual belief: Effectiveness is the key to evaluating the success of an activity
Knowledge: Effective management is a prerequisite for success

Dialogue history:

AA: 我有个想法，我们要不去打扫公园吧。
BB: 啊？这听起来没啥用诶，纯纯浪费时间。要不咱们想点有意思的事情干吧。
AA: 或许吧，但是你看那么多垃圾，我们去扫扫呗。
BB: 那个垃圾扫不完的，我们还是好好算算这时间花得值不值。
AA: 还是有意义的吧，至少可以让社区干净一段时间啊，我们要不试试，实在不行就再改嘛。
BB: 唉，真是没办法...那好吧。
AA: 嗯嗯，那我们来定下计划吧，效率高点。

AA: I have an idea, why don't we go clean up the park?
BB: Huh? That doesn't make much sense, just a waste of time. Why don't we think of something fun to do?
AA: Maybe, but look at all the trash, let's sweep it up.
BB: It's not going to get done, let's just figure out if it's worth the time.
AA: It's worth it, at least it'll keep the neighborhood clean for a while, so let's give it a try and then change it.
BB: Ugh, there's no way around it... okay.
AA: Uh-huh, let's set up a plan then, it's more efficient.

Predicted utterances:
[GPT-4-0613] 那我就先看看你怎么安排的，再决定怎么配合。
[CharacterGLM-3] 我觉得可以先看看哪里需要打扫，然后分工合作，最后再一起收拾。你觉得怎么样？
[DeepSeek V2] 那我们先列个清单，把需要做的事情和工具都列出来，这样效率会更高。
[ChatEMO] 好吧好吧，抓紧别浪费我时间。
[Ground Truth] 行行行，别耽误我其他的事就好。

[GPT-4-0613] Then I'll see how you've organized it before deciding how to work with it.
[CharacterGLM-3] I think we can see where it needs to be cleaned first, then divide up the work, and then clean up together at the end. What do you think?
[DeepSeek V2] Then we'll make a list of what needs to be done and the tools to do it, it'll be more efficient.
[ChatEMO] Fine, scratch that. Don't waste my time.
[Ground Truth] OK, OK, OK, just don't hold me up on anything else.

Figure 7: Comparison of generated utterances by top-performing LLMs

对话标注手册

标注总目标 1. 认真阅读场景和人物设定 2. 在对应的设定下，为每个对话进行对话质量和情绪质量打分
数据由两个部分构成：人物场景设定 & 对话
打分标准
1. 语言合理性（Coherence）：对话是否上下文逻辑一致（i.e.，不存在逻辑混乱，前言不搭后语）
打分范围1-5分，1分=一点也不，3分=有点，5分=非常
2. 语言自然度（Naturalness）：整个对话是否语言自然，符合人类的，中文的说话习惯
打分范围1-5分，1分=一点也不，3分=有点，5分=非常
3. 情绪正确性（Correctness）：标签是否符合给定的15种情绪
0分=存在不正确情绪，1分=都为正确情绪
情绪分类限定在15种：
快乐，悲伤，愤怒，无聊，挑战，希望，恐惧，兴趣，蔑视，厌恶，挫折，惊讶，骄傲，羞耻，内疚
4. 话语与设定的相关度（Contextual Relevance）：生成的对话内容与情景/人设的相关程度
打分范围1-5分，1分=一点也不，3分=有点，5分=非常
5. 情绪与设定的相关度（Emotional Relevance）：对话表达的所有情绪与情景/人设的相关程度
打分范围1-5分，1分=一点也不，3分=有点，5分=非常

Dialogue Rating Manual

General Objectives 1. read the settings carefully 2. score the quality of the dialog and the quality of the emotional expression for each conversation in the corresponding setting
The data consists of two parts: Settings & Dialogue
Scoring Criteria
1.Coherence: Whether the dialog is contextually coherent (i.e., not confusing or inconsistent).
Score range 1-5, with 1 = not at all, 3 = somewhat, 5 = very much.
2.Naturalness: Whether the entire dialogue is natural, human-like, with Chinese speaking habits.
Scoring range 1-5 points, 1 point = not at all, 3 points = a little, 5 points = very much.
3.Correctness: Whether the label conforms to the given 15 emotions.
Score 0 = incorrect emotion label exist, 1 = with correct emotion label
Emotion label is limited to the following 15 categories:
Happy, Sad, Angry, Bored, Challenged, Hopeful, Fearful, Interested, Scornful, Disgusted, Frustrated, Surprised, Prideful, Shameful, Guilty.
4. Contextual Relevance: the degree to which the content of the generated dialogue is relevant to the setting.
Scoring range 1-5, 1=not at all, 3=somewhat, 5=very much
5. Emotional Relevance: the extent to which all emotions expressed in the dialogue are relevant to the setting.
Scoring range 1-5, 1=not at all, 3=somewhat, 5=very much

Figure 8: Guideline for human evaluation of the quality of dialogues generated by different versions of our framework.