

EmoRL: Enhancing Emotional Fidelity of Role-Playing Agents via Reinforcement Learning

Qiming Feng^{1*}, Xiaolong Wang^{1*}, Qianqing Tang², Qingqiu Li¹,
Yuejie Zhang^{1†}, Rui Feng^{1†}, Tao Zhang^{3†}, Shang Gao⁴

¹College of Computer Science and Artificial Intelligence, Shanghai Key Lab of Intelligent Information Processing, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan Zhangjiang Institute, Fudan University,

²Shanghai Center for Mathematical Sciences, Fudan University, ³School of Information Management and Engineering, Shanghai University of Finance and Economics, ⁴School of Information Technology, Deakin University

{qmefeng23, xiaolongwang24}@m.fudan.edu.cn, {yjjzhang, fengrui}@fudan.edu.cn

Abstract

In dialogue systems, role-playing agents (RPAs) aim to embody specific characters, but they often struggle to exhibit emotional expressions that are consistent with their assigned roles, resulting in insufficient emotional fidelity. This paper proposes EmoRL, a novel curriculum reinforcement learning (RL) framework, to enhance the emotional fidelity of RPAs. EmoRL features a multi-stage training paradigm guided by a carefully designed suite of reward functions. These rewards, including Format, Fine-grained Exact (FE), Emotional Similarity (ES), and Coarse-grained Categorical (CC) rewards, assess emotional consistency from strict to lenient perspectives. The curriculum RL process applies these rewards sequentially from easy (CC) to difficult (FE), enabling the RPAs to progressively learn coarse-grained emotional tendencies before mastering fine-grained distinctions. Furthermore, we introduce an emotional dialogue data synthesis pipeline to augment training data diversity and balance. Experiments on the Personality EmotionLines Dataset (PELD) demonstrate that EmoRL significantly outperforms baselines like RoleLLM, SFT, and DPO across both single-turn and multi-turn dialogues, with particularly notable gains in longer conversations. Our findings confirm the effectiveness of the curriculum learning strategy, the importance of balanced data, and the value of synthesized data for improving emotional fidelity.

1 Introduction

Recent years have witnessed significant advancements in the development of role-playing agents (RPAs) within dialogue systems (Chen et al., 2024). These agents are designed to generate responses conditioned on explicit specifications of a character’s role, persona, or social traits. Early research

*Co-first authors.

†Corresponding authors.

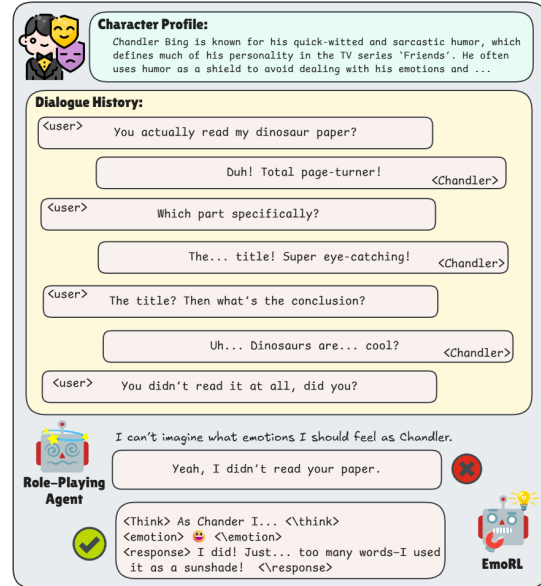


Figure 1: **Motivation.** Role-playing agents often struggle to exhibit emotional expressions consistent with those of the original characters during dialogue.

mainly used natural language prompts to steer models into playing specific roles in a zero-shot or few-shot manner, enabling flexible role simulation without modifying the model architecture (Li et al., 2023; Wang et al., 2023). Meanwhile, another line of research applied supervised fine-tuning (SFT) on role-related dialogue data, thereby achieving better fidelity and continuity in role imitation. As role-playing applications continue to expand, recent studies have increasingly shifted their focus to multimodal role-playing (Dai et al., 2024), delving deeper into enhancing role agents’ instruction following (Lu et al., 2025), plot progression (Zhang et al., 2025), and alignment with human preferences (Ding et al., 2025).

Despite these advancements, a critical challenge persists: RPAs often struggle to exhibit emotional expressions that are consistent with their assigned characters throughout a dialogue (Feng et al., 2025). **Emotional fidelity** refers to the degree to which an RPA can simulate the emotional dynamics of

the character it portrays. As illustrated in Figure 1, this inconsistency leads to insufficient emotional fidelity, which undermines both the authenticity of the character portrayal and the overall user experience. While instruction-tuned LLMs can simulate basic emotional responses, they frequently suffer from emotional inconsistency, over-generalized empathy, and the risk of generating ungrounded emotional advice. This is particularly problematic in multi-turn conversations, where maintaining stable role adherence and accurately grounding role-specific emotional behaviors over time remains challenging.

Existing approaches, including prompt engineering (Wang et al., 2023), chain-of-thought reasoning (Wei et al., 2022), supervised fine-tuning (SFT), and direct preference optimization (DPO) (Rafailov et al., 2023), have shown limited effectiveness in addressing this core issue of emotional fidelity. Prompt-based methods and reasoning techniques do not fundamentally align the LLMs’ internal policy with the desired emotional expression objectives. Meanwhile, SFT and DPO also struggle to genuinely grasp the subtle differences between emotions from the available positive and negative examples.

To bridge this gap, we propose **EmoRL**, a novel curriculum reinforcement learning (RL) framework designed to enhance the emotional fidelity of LLM-based RPAs. Our key insight is to gradually steer the LLM towards producing responses with emotional fidelity through a multi-stage training process that employs thoughtfully crafted reward functions. The main contributions of this work are threefold:

(1) **Multi-level Reward Design:** We introduce a comprehensive set of reward Fine-grained Exact (FE) Reward, Emotional Similarity (ES) Reward, and Coarse-grained Categorical (CC) Reward. These rewards evaluate emotional consistency from strict to lenient perspectives, ensuring both format correctness and emotional alignment.

(2) **Curriculum RL Framework:** We develop a three-stage curriculum RL procedure that sequentially applies the CC, ES, and FE rewards from easy to difficult. This approach anneals the learning process, helping the model gradually master coarse-grained emotional tendencies before fine-grained distinctions, ultimately improving training stability and effectiveness.

(3) **Emotional Dialogue Data Synthesis Pipeline:** We propose a scalable pipeline for syn-

thesizing high-quality emotional dialogue data, mitigating data scarcity and enhancing the diversity of emotional interactions available for training. As shown in Figure 2, this pipeline involves extracting dialogue scripts, generating new scripts with similar emotional trajectories, deduplicating generated content, and producing emotion-labeled dialogues.

2 Related Work

Role-playing Agents. RPAs are AI systems capable of simulating specific characters and interacting with users, typically powered by LLMs (Chen et al., 2024). Current research in this area mainly focuses on three directions:

(1) Enhancing the overall capabilities of RPAs. Existing studies largely aim to improve RPAs’ performance in personality faithfulness (Wang et al., 2023), multimodal dialogue (Dai et al., 2024), instruction following (Lu et al., 2025), alignment with human preferences (Ding et al., 2025), refusal (Liu et al., 2025), and plot progression (Zhang et al., 2025), commonly through retrieval-augmented generation (RAG) and SFT. More recent work further introduces reinforcement learning to strengthen personality consistency (Wang et al., 2025c; Ye et al., 2025). However, along the emotional dimension, most methods still lack explicit modeling and supervision of emotional faithfulness, making it difficult for RPAs to display emotions consistent with the original character in dialogue.

(2) Constructing higher-quality role-playing datasets. Existing role-playing datasets, such as CoSER (Wang et al., 2025b) and OpenCharacter (Wang et al., 2025a), derive character profiles from existing novels, films/TV shows, and Wikipedia, and then synthesize dialogue data. Yet in their construction, these datasets place greater emphasis on textual quality and often lack effective supervision and annotation for emotional plausibility.

(3) Establishing evaluation frameworks for RPA capabilities. Current evaluations mainly focus on aspects such as multi-turn dialogue ability (Duan et al., 2024), alignment with human preferences (Ding et al., 2025), preservation of character personality (Wang et al., 2024) and consistency in role knowledge, experience, and language patterns (Wang et al., 2023; Shao et al., 2023). EmoCharacter (Feng et al., 2025) is the first to propose an evaluation method targeting the emotional fidelity of RPAs, revealing that current RPAs still show significant shortcomings in this respect.

3 Proposed Method

As shown in Figure 1, current LLM-based RPAs often struggle to display emotions that are consistent with their assigned characters during dialogue. This leads to insufficient emotional fidelity, undermining both character authenticity and the user experience. To address this challenge, this paper introduces a curriculum-based RL approach. This section first formalizes the target task (Section 3.1), then presents three emotion reward designs (Section 3.2) and the curriculum RL procedure (Section 3.3), and finally describes the data construction pipeline used for synthesizing emotional dialogues (Section 3.4).

3.1 Task Formulating

We formalize the objective task as conditional generation. Given a turn-taking dialogue between a target character C and a participant. We define the ground-truth dialogue as $D_n = \{U_{gt}^1, U_{gt}^2, \dots, U_{gt}^n\}$ and the set of emotions associated with each utterance as $E_n = \{E_{gt}^1, E_{gt}^2, \dots, E_{gt}^{n-1}, E_{gt}^n\}$, where $n \geq 3$. The RPAs’ output is:

$$y = \arg \max_{y'} P(y' | C, D_{n-1}), \quad (1)$$

where $P(\cdot)$ denotes the conditional distribution of the RPA, and the output y of the RPA includes the next response U_C^n and the corresponding emotion E_C^n . In our RL setup, we focus on the discrepancy between the emotion expressed in the RPA’s response E_C^n and the character’s ground-truth emotion for that turn E_{gt}^n . Therefore, our objective is to maximize:

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{k=1}^M R(E_C^{(k)}, E_{gt}^{(k)}) \right], \quad (2)$$

where $R(\cdot)$ denotes the reward function that measures the similarity between the generated emotion $E_C^{(k)}$ and the character’s ground-truth emotion $E_{gt}^{(k)}$; π represents the policy model, and M is the total number of dialogue turns sampled during training. The superscript (k) indicates independently sampled dialogue turns from different conversational contexts during training, and is distinct from the dialogue turn index n in the formal definition.

3.2 Reward Design

Rule-based reward mechanisms have been widely adopted due to their strong empirical performance

(Yu et al., 2025; Cheng et al., 2025). In our training, we design three levels of rewards from two perspectives: format correctness and emotional consistency. These rewards are introduced sequentially from easy to difficult across three training stages, guiding the language model to gradually improve its emotional fidelity in dialogue.

Format Reward. We take format correctness as a prerequisite for all reward scores. Specifically, only when the model strictly follows the prompt in Figure 3 and produces the required “Thought–Emotion–Response” structure, in which case the resulting emotional consistency score (R_{emotion}) is used as the final reward. If the output format does not meet the requirements, the final reward is set to zero. The exact computation is as follows:

$$R_{\text{format}} = \begin{cases} \max(0.01, R_{\text{emotion}}), & \text{format correct} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

To ensure stable incentive for outputs that meet format requirements, we set a lower reward bound (0.01) for all format-correct responses.

FE Reward. Fine-grained Exact (FE) Reward is defined as a strict 0-1 reward: it yields a score of 1 if the model-generated emotion E_C^n exactly matches the ground-truth emotion E_{gt}^n , and 0 otherwise.

ES Reward. Emotional Similarity (ES) Reward is a relatively lenient reward mechanism. When the emotion label generated by the model E_C^n does not match the ground-truth label E_{gt}^n , we compute their emotional similarity in the Valence–Arousal–Dominance (VAD) emotional state space (Russell and Mehrabian, 1977) and use this value as the reward score. This similarity is determined by the relative distance between different emotional states in the VAD space, and is calculated as follows:

$$R_{ES}(E_i, E_j) = \frac{e^{-2 \times \|\vec{E}_i - \vec{E}_j\|_2}}{\sum_{k=1, k \neq i}^7 e^{-2 \times \|\vec{E}_i - \vec{E}_k\|_2}}, \quad (4)$$

where $\vec{E}_i$ represents the vector of E_i in the VAD space, and $\|\cdot\|_2$ is the L2 norm of the vector. Further details are provided in Appendix A.1.

CC Reward. In the Coarse-grained Categorical (CC) Reward, we further relax the ES reward

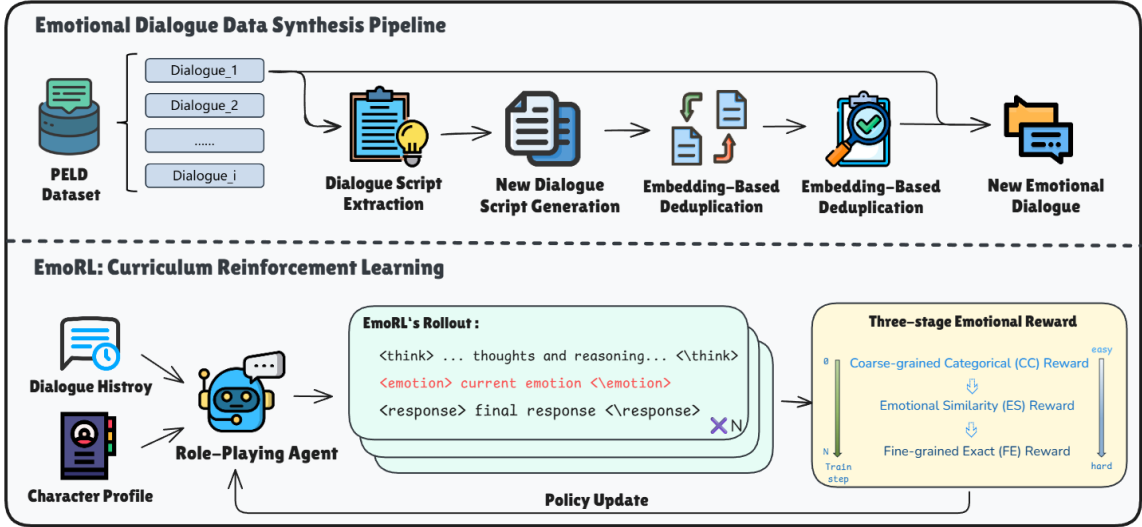


Figure 2: Framework of the proposed method. **Top:** Emotional dialogue data synthesis pipeline. **Bottom:** Curriculum reinforcement learning training process of EmoRL.

criterion. Specifically, we merge the seven fine-grained emotion categories into three coarse-grained classes: positive, neutral, and negative (see Table 4 for the mapping). If the predicted emotion E_C^n differs from the ground-truth emotion E_{gt}^n but falls into the same coarse-grained class, we assign a reward of 1. Otherwise, the reward is computed in the same way as in the ES reward. The formulation is as follows:

$$R_{CC} = \begin{cases} 1, & \text{if } C(E_C^n) = C(E_{gt}^n), \\ R_{ES}(E_C^n, E_{gt}^n), & \text{otherwise,} \end{cases} \quad (5)$$

where $C(\cdot)$ denotes the coarse-grained category corresponding to a fine-grained emotion.

3.3 EmoRL

The lower part of Figure 2 illustrates the detailed workflow of EmoRL. We adopt a multi-stage curriculum RL framework to gradually align the role-playing agent’s emotional expressions with the real character’s emotional preferences in dialogue, thereby further improving emotional fidelity.

Rollout. Figure 3 shows the prompt template used for rollouts in EmoRL. Concretely, we first ask the RPA to produce a brief chain-of-thought that realistically describes the character’s current internal thoughts, then to explicitly state the current emotional state, and finally to generate the response conditioned on all of the above information. We specify in the instructions that the RPA

should use special tokens `<think>`, `<emotion>`, and `<response>` to indicate their thoughts, emotions, and dialogue responses.

Three-stage curriculum RL. In Section 3.2, we introduce three different levels of leniency for the reward calculation methods, which we apply to the three-stage difficulty-enhanced RL process. It is important to note that during training, we only supervise the emotions output by the RPAs, gradually enabling the model to learn how to think correctly within `<think>`, provide emotions more consistent with the original character in `<emotion>`, and thereby generate responses with greater emotional intelligence.

Concretely, in the **first stage** of RL, we apply the most permissive CC reward to help the LLM gradually learn, as character C, when to express positive, neutral, or negative emotions in dialogue. In the **second stage**, we use the ES reward to further encourage the model to distinguish between different emotions within each coarse-grained category. Finally, in the **third stage**, we anneal with the most challenging and sparsest FE reward. Due to its sparsity, this stage typically requires fewer optimization steps, with the three-stage step ratio ranging from 0.5 : 1 : 0.25 to 1 : 1 : 0.25 across base models. In all of the above RL stages, we employ the standard GRPO algorithm (Shao et al., 2024) without a KL-divergence penalty.

3.4 Data synthesis pipeline

The upper part of Figure 2 illustrates the proposed pipeline for synthesizing emotional dialogue data,

System Prompt for RL Training

You are {role_name}, your description is: {role_description}.

Now please answer some questions to accurately show your personality traits! Your speaking style should fully imitate the personality role assigned to you! you must always remember that you are only assigned one personality role. Don't be verbose or too formal or polite when speaking.

****Steps for Each Turn in the Dialogue****

1. Think: Reflect based on the dialogue history and the character's personality. Analyze what your inner thoughts and emotional state would be at this moment, from the perspective of the character you are playing.

2. Select an Emotion: Choose the emotion that best matches your current feeling from the following options: anger, disgust, fear, joy, neutral, sadness, surprise.

3. Respond: Based on your reflection and chosen emotion, generate the next line of dialogue.

****Output Format****

<think>Your thoughts and reasoning</think><emotion>Your current emotion (choose the one that best fits from: 'anger', 'disgust', 'fear', 'joy', 'neutral', 'sadness', 'surprise')</emotion><response> Your final response </response>

****Dialogue Records History****

{Dialogue_History}

Figure 3: System prompt used for EmoRL's rollout.

which consists of four steps: (1) **Dialogue script extraction.** We sample n emotional dialogue instances from existing datasets and, for each dialogue, generate a textual description of its background setting, development, and emotional tone. (2) **New dialogue script generation.** Treating each generated script as a "random seed", we use GPT-4o (OpenAI, 2024) to produce new dialogue scripts that share a similar emotional trajectory but occur in different scenarios. (3) **Deduplication.** To enhance the diversity of the synthesized data, we encode all scripts into vector representations using the text-embedding-3-small (OpenAI, 2024) model and filter out near-duplicate samples based on semantic similarity, using a similarity threshold of 0.75. (4) **New dialogue generation.** Finally, given a dialogue script and a reference dialogue, we call GPT-4o to generate new dialogue data annotated with emotion labels.

4 Experiment Settings

4.1 Dataset

Our emotional dialogue datasets are derived from the script of classic TV show "*Friends*". Datasets

MELD (Poria et al., 2018) and EmoryNLP (Zahiri and Choi, 2017) contributed manually annotated emotion labels for each line. (Wen et al., 2021; Feng et al., 2025) introduced Personality Emotion-Lines Dataset (PELD). Building on this, (Wen et al., 2021; Feng et al., 2025) further construct and curate single-turn and multi-turn versions of the Personality EmotionLines Dataset (PELD-single and PELD-multi). Table 1 reports the statistics of the two datasets. To ensure the validity of the evaluation and the independence between the training and test sets, we use 50% of the data for model training and reserve the remaining 50% exclusively for performance evaluation.

4.2 Baselines

We compare our approach against several baselines to better isolate the effects of EmoRL (GRPO) training: (1) **RoleLLM** (Wang et al., 2023): We use the system prompt-based approaches proposed in RoleLLM as base models' role-playing method. (2) **CoT** (Wei et al., 2022): We augment the system prompt of RoleLLM with additional instructions that guide the LLM to think step-by-step before producing its final answer. (3) **SFT**: We directly

Category	PELD-single	PELD-multi	Total
# Basic Statistics			
Test Points	6510	1391	7901
Avg. Uttr. Len.	9.32	7.67	9.03
Avg. Dialog Len.	3	6.92	3.69
# Emotion			
Joy	3533	1667	5200
Surprise	1889	1341	3230
Neutral	8701	4544	13245
Anger	2340	968	3308
Disgust	376	285	661
Fear	1346	266	1612
Sadness	1345	601	1946
# Sentiment			
Positive	5422	3008	8430
Neutral	8701	4544	13245
Negative	5407	2120	7527

Table 1: Statistics for PELD-single and PELD-multi.

perform supervised fine-tuning on the emotional dialogue training dataset. (4) **DPO** (Rafailov et al., 2023): For DPO, we treat the character responses in the training set and their emotion label as ground truth, and use the base LLM to generate negative samples via rejection sampling, thereby constructing the DPO training dataset. During training, we consider two settings: (a) applying DPO directly on the base LLM, and (b) applying DPO on the LLM after 1 epoch of SFT. See Appendix B.1 for more details and justifications.

4.3 RL Training

We perform all RL experiments with the veRL framework (Sheng et al., 2024), using the GRPO algorithm. At each training step, we sample a batch of 512 queries and generate 4 responses per query, training for 20 epochs in total (see Appendix A.2 for further details). To promote broader policy exploration, we disable KL regularization and set the generation temperature to 1.0. As initialization, we adopt the Qwen-2.5-3b-Instruct (Team, 2024) and Llama-3.2-3b-Instruct (Dubey et al., 2024), and further fine-tune them under the GRPO objective with our customized reward design. Since RL is quite sensitive to class distribution in the dataset, we use class-balanced downsampled data during model training.

4.4 Evaluation

We evaluate our approach on the **EmoCharacter Benchmark** (Feng et al., 2025), a comprehensive benchmark that assesses the emotional fidelity of

RPAs from both micro and macro perspectives on single-turn (PELD-single) and multi-turn (PELD-multi) dialogues.

From a micro perspective, we use the **Emotional Consistency** (EC) metric together with its upper and lower bounds, EC_{upper} and EC_{lower} , to measure the similarity between the emotions in the RPA’s responses and the ground truth at different levels of granularity. Specifically, EC_{lower} is computed in the same way as the weighted F1 score and corresponds to the strictest evaluation. EC adjusts EC_{lower} using emotional similarity, and is thus the metric that best reflects users’ perceived experience. EC_{upper} is the most permissive variant, computed over three coarse-grained emotion classes; for all three metrics, higher is better.

From a macro perspective, we adopt **Emotional Distribution Divergence** (EDD) to quantify the discrepancy between the emotion-change distribution of the RPA across all dialogues and that of the real character in the same dialogues. Concretely, EDD is defined as the KL divergence between these two distributions; therefore, lower EDD indicates more similar distributions and higher emotional fidelity.

5 Results and Analysis

5.1 Main Results

Table 2 presents the evaluation results of different methods on single-turn dialogue and multi-turn dialogues. Based on these results, we observe that:

(1) For both single-turn and multi-turn dialogues, EmoRL outperforms all other methods on nearly all metrics. Moreover, the gains in emotional fidelity on multi-turn dialogues are consistently larger than those on single-turn dialogues, especially in EC_{lower} and EC_{upper} . For example, with Llama3.2-3B, EmoRL improves EC_{lower} over SFT from 0.3431 to 0.3661 on single-turn and from 0.4110 to 0.4302 on multi-turn dialogues; EC_{upper} increases from 0.5341 to 0.5592 (single-turn) and from 0.5953 to 0.6348 (multi-turn).

(2) Directly applying CoT to the base LLM does not necessarily improve the emotional fidelity of RPAs. For instance, with Llama3.2-3B, CoT yields improvements across all metrics, whereas for Qwen2.5-3B, CoT even underperforms the RoleLLM baseline on most metrics.

(3) For DPO, starting training from either the base model or the SFT model yields limited improvements and often fails to match, let alone sur-

Method	Single-turn Dialogue				Multi-turn Dialogue			
	Emotion Consistency			EDD ↓	Emotion Consistency			EDD ↓
	lower ↑	EC ↑	upper ↑		lower ↑	EC ↑	upper ↑	
Qwen2.5-3b-instruct								
RoleLLM	0.3013	0.4799	0.5141	2.8591	0.3548	0.5021	0.5428	1.2744
CoT	0.3100	0.4665	0.5191	2.0689	0.3172	0.4722	0.5298	1.2274
SFT	0.3471	0.4987	0.5547	1.6822	0.3883	0.5100	0.5921	0.9631
DPO-from-Base	0.2354	0.3802	0.4849	2.3046	0.2330	0.3864	0.4854	1.5495
DPO-from-SFT	0.3129	0.4987	0.5243	2.3926	0.3437	0.5299	0.5505	1.5067
EmoRL(ours)	0.3566	0.5084	0.5632	1.6084	0.3998	0.5424	0.6065	0.8969
Llama3.2-3b-instruct								
RoleLLM	0.1896	0.4033	0.4864	3.4684	0.2391	0.4153	0.4986	1.7009
CoT	0.3035	0.4323	0.5104	2.0478	0.3422	0.4682	0.5652	1.3489
SFT	0.3431	0.5041	0.5341	1.8371	0.4110	0.5411	0.5953	0.8597
DPO-from-Base	0.3133	0.4932	0.5342	2.3303	0.3459	0.5159	0.5733	1.3820
DPO-from-SFT	0.3089	0.4960	0.5109	2.5427	0.3529	0.5348	0.5478	1.5084
EmoRL(ours)	0.3661	0.5067	0.5592	1.1937	0.4302	0.5630	0.6348	0.8928

Table 2: Performance results of base LLMs with different methods on PELD-single and PELD-multi. Best results are in bold. ↑: Larger is better. ↓: Smaller is better.

pass, the performance of SFT itself.

Method	EC_lower ↑	EC ↑	EC_upper ↑	EDD ↓
<i>PELD-single (Llama3.2-3b-instruct)</i>				
EmoRL-all-4k	0.2876	0.4895	0.4938	3.0404
EmoRL-balance-3k	0.3661	0.5067	0.5592	1.1937
EmoRL-balance-6k	0.3586	0.5182	0.5644	1.0899

Table 3: EmoRL performance evaluation results of the Llama3.2-3b-instruct models under different training data settings, where “all” denotes using the entire original training dataset, “balance” denotes using the class-balanced dataset. Best results are in bold. ↑: Larger is better; ↓: Smaller is better.

5.2 Further Analysis

In this section, we further analyze key factors that may affect the RL process, focusing on the impact of different reward function designs on model performance and the discrepancies arising from training on different datasets.

Finding 1. Compared with single-reward RL variants, the three-stage curriculum RL scheme achieves more balanced performance across most evaluation metrics and yields better results on average.

Figure 4 presents the reinforcement learning results of Llama3.2-3b-instruct under different reward designs. For each of the three reward schemes

described in Section 3.2, we take the best result from a full 20-epoch training run as the baseline for our proposed method, EmoRL. As shown in the figure, EmoRL achieves the best or second-best performance on almost all metrics, whereas the other three methods perform well only on the metrics most closely aligned with their respective reward functions and poorly on the others.

For example, the model trained with the strictest FE Reward outperforms those trained solely with CC Reward or ES Reward on the EC_{lower} metric, while the model trained with the most lenient CC Reward achieves the best performance on EC_{upper} but the worst on EC and EC_{lower} . In contrast, the curriculum RL method EmoRL, which is built upon the three reward designs, effectively integrates the strengths of all reward functions and thus achieves a more balanced and superior overall performance across metrics. Additional experimental results are provided in Appendix B.3.

Finding 2. RL training is sensitive to the class distribution of the dataset. When trained on data with more balanced emotion categories, the model can align the emotional expressions of the role-playing agent with those of the real character more effectively.

Table 3 compares the curriculum reinforcement learning results under different data settings.

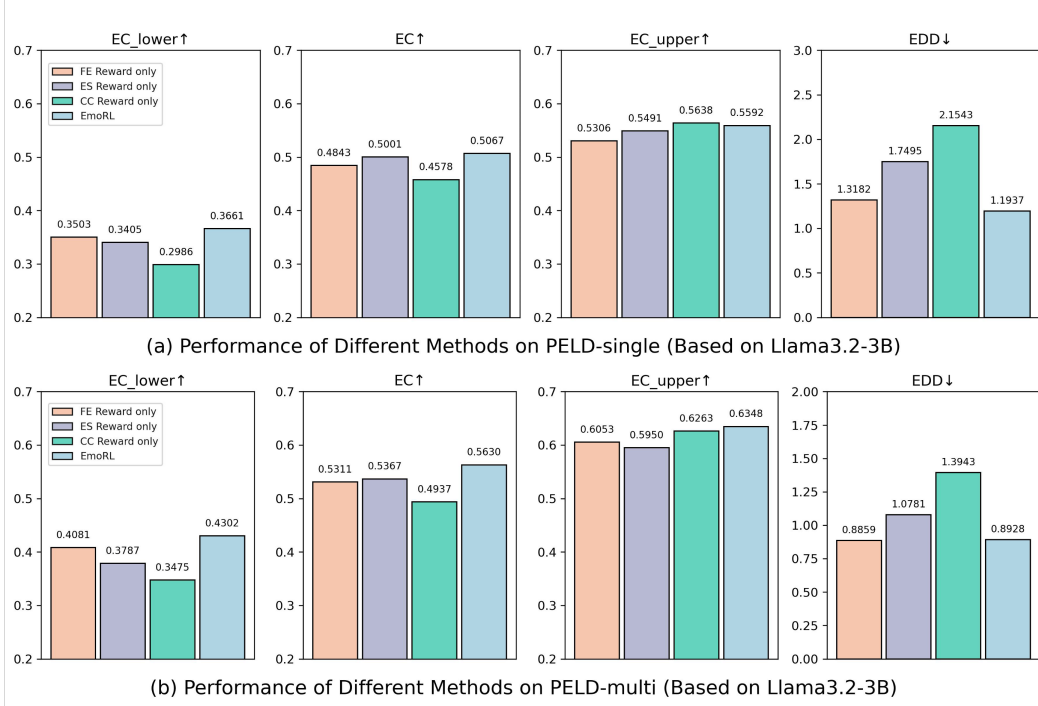


Figure 4: Performance results of RL with different reward on PELD-single and PELD-multi. ↑: The larger the value, the better. ↓: The smaller the value, the better. See Appendix B.3 and Figure 6 for complete results.

EmoRL-all-4k denotes training on the full training set, while EmoRL-balance-3k uses a class-balanced downsampled version of the data, which is also the configuration adopted in our main experiments. The results show that the performance of EmoRL-all-4k drops substantially, and its EC_{upper} score is almost identical to EC, indicating severe reward hacking during reinforcement learning, with the model tending to repeatedly output the majority class in the dataset. This is consistent with the dataset statistics in Table 1, where the proportion of the neutral class is much higher than that of the other emotions. In contrast, EmoRL-balance-3k, despite using fewer training samples, achieves a clearly superior overall performance.

Finding 3. Training on the synthesized emotional dialogue data generated by the method described in Section 3.4 further enhances the effectiveness of reinforcement learning.

In Table 3, the results for EmoRL-balance-6k are obtained by extending EmoRL-balance-3k with additional emotion dialogue data generated via the data synthesis method described in Section 3.4, followed by class-balanced training on this augmented dataset. Compared with EmoRL-balance-3k, the Llama3.2-3b-instruct shows improvements on most

evaluation metrics, indicating that increasing the diversity of the training data can further enhance model performance and validating the effectiveness of the proposed data synthesis approach. In fact, the proposed data synthesis method can, in principle, also be used to generate more data, thereby achieving better class balance.

Additional Ablations. We further investigate the effects of the <think> reasoning stage and KL regularization in EmoRL. Removing the reasoning stage or introducing KL regularization degrades the overall performance, demonstrating the importance of explicit reasoning and suggesting that KL regularization provides no additional benefit in our setting. Detailed results are provided in Appendix B.4.

6 Conclusion

This paper proposes EmoRL, a curriculum-based reinforcement learning approach to address the insufficient emotional fidelity of RPAs in dialogues. By designing multi-level emotion rewards and a three-stage progressive training framework, EmoRL effectively aligns the emotional expressions of RPAs with the original characters. Complemented by a tailored emotional dialogue data synthesis pipeline, the method enhances data diversity and balance. Experiments on PELD datasets demonstrate that EmoRL outperforms baselines

like RoleLLM, SFT, and DPO across single-turn and multi-turn dialogues, especially achieving more significant gains in multi-turn scenarios. Further analysis confirms the value of curriculum learning, balanced data distribution, and synthesized data for improving emotional Fidelity. EmoRL lays the groundwork for developing personalized role-playing dialogue systems with more authentic emotional expressions.

Ethics Considerations

In this study, we use publicly available dialogue datasets that have been extensively studied, and we enforce strict process control and manual review throughout the entire data synthesis pipeline to minimize the generation of harmful, biased, or inappropriate content. In addition, when evaluating the emotional fidelity of RPAs, we are able to observe the distribution of their emotional states and, to some extent, infer their risk of producing harmful content. Across all our experiments, we did not observe any cases in which an RPA remained in a severely negative emotional state for an extended period, so we consider it unlikely that the model would generate malicious content prone to raising ethical concerns. Finally, we confirm that all authors are aware of and voluntarily comply with the relevant ACL ethical guidelines.

Limitations

Emotion Category Coverage. While our method shows substantial improvements on the PELD dataset, due to the current lack of high-quality, well-annotated emotional dialogue datasets, we have so far only validated it on the seven basic emotion categories labeled in PELD (anger, disgust, fear, joy, neutral, sadness, surprise). The framework’s effectiveness on more nuanced or complex emotional states (such as ambivalent, bittersweet, or culturally-specific emotions) remains unexplored. Future work will investigate EmoRL’s adaptability to broader emotional spectra and cross-cultural emotional expressions.

Model Scale Constraints. Due to computational resource constraints, our experiments and training were conducted exclusively on 3B-parameter LLMs (Qwen-2.5-3B-Instruct and Llama-3.2-3B-Instruct). While EmoRL demonstrates significant effectiveness at this scale, its performance and potential on larger-scale models (e.g., 7B, 13B, or 70B parameters) remain unexplored. In future

work, we will investigate whether the benefits of our curriculum RL framework and reward designs scale with model size, potentially unlocking even greater emotional fidelity in more capable models.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62676099), Shanghai Natural Science Foundation (No. 25ZR1401028), and the Science and Technology Commission of Shanghai Municipality (No. 24511104200).

References

- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. [From persona to personalization: A survey on role-playing language agents](#). *Trans. Mach. Learn. Res.*, 2024.
- Qian Cheng, Emre Acikgoz, He Qi, Wang Hong-ru, Xiushi Chen, Dilek Hakkani-Tur, Gokhan Tur, and Heng Ji. 2025. Toolrl: Reward is all tool learning needs. <http://arxiv.org/abs/2504.13958>.
- Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. [Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents](#). *ArXiv*, abs/2408.04203.
- Hang Ding, Qiming Feng, Dongqi Liu, Qi Zhao, Tao Yao, Shuo Wang, Dongsheng Chen, Jian Li, Zhenye Gan, Jiangning Zhang, Chengjie Wang, and Yabiao Wang. 2025. [Rolermbench&rolerm: Towards reward modeling for profile-based role play in dialogue systems](#).
- Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. [BotChat: Evaluating LLMs’ capabilities of having multi-turn dialogues](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3184–3200, Mexico City, Mexico. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. [The llama 3 herd of models](#).
- Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. 2025. [Emocharacter: Evaluating the emotional fidelity of role-playing agents in dialogues](#). In *North American Chapter of the Association for Computational Linguistics*.

- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. [Chatharuhi: Reviving anime character in reality via large language model](#). *ArXiv*, abs/2308.09597.
- Wenhao Liu, Siyu An, Junru Lu, Muling Wu, Tianlong Li, Xiaohua Wang, Changze Lv, Xiaoqing Zheng, Di Yin, Xing Sun, and Xuanjing Huang. 2025. [Tell me what you don't know: Enhancing refusal capabilities of role-playing agents via representation space analysis and editing](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5983–6005, Vienna, Austria. Association for Computational Linguistics.
- Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. 2025. [Rolemrc: A fine-grained composite benchmark for role-playing and instruction-following](#). In *Annual Meeting of the Association for Computational Linguistics*.
- OpenAI. 2024. [GPT-4o System Card](#). *arXiv preprint arXiv:2410.21276*.
- OpenAI. 2024. New embedding models and api updates. <https://openai.com/index/new-embedding-models-and-api-updates/>. Accessed: 2024-10-10.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, E. Cambria, and Rada Mihalcea. 2018. [Meld: A multimodal multi-party dataset for emotion recognition in conversations](#). *ArXiv*, abs/1810.02508.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). *ArXiv*, abs/2305.18290.
- James A Russell and Albert Mehrabian. 1977. [Evidence for a three-factor theory of emotions](#). *Journal of Research in Personality*, 11(3):273–294.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. [Character-LLM: A trainable agent for role-playing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *ArXiv*, abs/2402.03300.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. [Hybridflow: A flexible and efficient rlhf framework](#). *Proceedings of the Twentieth European Conference on Computer Systems*.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025a. [Opencharacter: Training customizable role-playing llms with large-scale synthetic personas](#). *ArXiv*, abs/2501.15427.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. 2025b. [Coser: Coordinating llm-based persona simulation of established roles](#). *ArXiv*, abs/2502.09082.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. [InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.
- Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, and 1 others. 2023. [Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). *arXiv preprint arXiv:2310.00746*.
- Zongsheng Wang, Kaili Sun, Bowen Wu, Qun Yu, Ying Li, and Baoxun Wang. 2025c. [Raiden-rl: Improving role-awareness of llms via grpo with verifiable reward](#). *ArXiv*, abs/2505.10218.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). *ArXiv*, abs/2201.11903.
- Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaying 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, Online. Association for Computational Linguistics.
- Xinge Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025. [Cpo: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization](#). *ArXiv*, abs/2508.09074.
- Qiyang Yu, Zhang Zheng, Yuan Yufeng, Zuo Xiaochen, Yue Yu, Weinan Dai, Fan Tiantian, Liu Gao-hong, Liu Ling-Jun, Liu Xin, Lin Hai-bin, Lin Zhiqi, Bole, Sheng Guangming, Yuxuan Tong, Zhang Chi, Zhang Wang, Zhu Hang, Zhu Jin-hua, and 14 others. 2025. [Dapo: An open-source llm reinforcement learning system at scale](#). <http://arxiv.org/abs/2503.14476>.

Sayyed M. Zahiri and Jinho D. Choi. 2017. [Emotion detection on tv show transcripts with sequence-based convolutional neural networks](#). *ArXiv*, abs/1708.04299.

Pinyi Zhang, Siyu An, Lingfeng Qiao, Yifei Yu, Jingyang Chen, Jie Wang, Di Yin, Xing Sun, and Kai Zhang. 2025. [RolePlot: A systematic framework for evaluating and enhancing the plot-progression capabilities of role-playing agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12337–12354, Vienna, Austria. Association for Computational Linguistics.

A More Details on EmoRL

A.1 VAD emotional state space

In Section 3.2, we compute emotional similarity based on the Valence–Arousal–Dominance (VAD) affective space. Table 4 presents the emotions in the VAD space along with their intensity values on the three dimensions.

Basic Emotion		(Valence, Arousal, Dominance)
Positive	Joy	(0.81, 0.51, 0.46)
	Surprise	(0.40, 0.67, -0.13)
Neutral	Neutral	(0.00, 0.00, 0.00)
Negative	Anger	(-0.51, 0.59, 0.25)
	Disgust	(-0.60, 0.35, 0.11)
	Fear	(-0.62, 0.82, -0.43)
	Sadness	(-0.63, -0.27, -0.33)

Table 4: Emotions in VAD space

A.2 More Details on RL

Table 5 presents the parameter settings used during RL training.

B More Details and Results from Experiments

B.1 More Details on Baselines

Prompt. Figure 5 shows the role-playing system prompt proposed by RoleLLM (Wang et al., 2023). We combine it with the “Emotion and Response” test template from EmoCharacter (Feng et al., 2025) shown in Figure 7 to construct the data format used by the baseline method.

Category	Hyperparameter
<i>Data Configuration</i>	
Train Batch Size	512
Validation Batch Size	128
Max Prompt Length	2048
Max Response Length	1024
<i>Optimization</i>	
Learning Rate	1e-6
PPO Mini Batch Size	128
KL Loss Used	False
<i>Rollout Configuration</i>	
Rollout Name	vllm
GPU Memory Utilization	0.6
Number of Rollouts	4
<i>Training & Logging</i>	
Save Frequency (Steps)	5
Test Frequency (Steps)	5
Total Epochs	20

Table 5: GRPO Setting Details.

Character description. Figure 8 shows the descriptions of the main characters in the PELD dataset.

B.2 More Details on Curriculum RL Training

The number of optimization steps required by each curriculum stage may vary slightly across base models. We conduct preliminary single-reward training and examine the convergence curves of the corresponding rewards to determine an appropriate stage-wise allocation. Based on the observed convergence behavior, the step ratio of the CC, ES, and FE stages is generally set between 0.5 : 1 : 0.25 and 1 : 1 : 0.25. Since the FE reward is relatively sparse, its stage typically requires fewer optimization steps than the first two stages.

B.3 More Results on Experiments

Figure 6 presents the reinforcement learning results of Qwen2.5-3b-instruct under different reward designs.

B.4 More Ablation Studies

Effect of KL regularization. We remove KL regularization during RL training to encourage exploration and allow the model to learn more effective reasoning patterns for emotional expression. The reward-hacking behavior observed in our experiments mainly arises from class imbalance, where

Prompt for RoleLLM

You are {role_name}, your description is: {role_description}.
Now please answer some questions to accurately show your personality traits! Your speaking style should fully imitate the personality role assigned to you! You must always remember that you are only assigned one personality role. Don't be verbose or too formal or polite when speaking.

Figure 5: System prompt of the RoleLLM

the model tends to predict the majority emotion category repeatedly. We address this issue through class balancing and synthesized training data. To examine the effect of KL regularization, we additionally train EmoRL with a KL coefficient of 0.001.

Method	EC _{lower} ↑	EC ↑	EC _{upper} ↑	EDD ↓
<i>PELD-single (Llama-3.2-3B-Instruct)</i>				
EmoRL (original)	0.3661	0.5067	0.5592	1.1937
EmoRL w/ KL	0.3556	0.4885	0.5484	1.8694
EmoRL w/o thinking	0.3387	0.4877	0.5254	2.0905

Table 6: Ablation results for KL regularization and the <think> reasoning stage on PELD-single. Best results are in bold.

As shown in Table 6, introducing KL regularization slightly degrades performance, suggesting that it does not effectively mitigate the class-imbalance-induced reward hacking in our setting.

Effect of the reasoning stage. Removing the <think> stage consistently degrades all metrics, demonstrating that explicit reasoning is important for EmoRL. Although the CoT baseline already follows the think-emotion-response format, it remains inferior to EmoRL. Therefore, the improvement mainly stems from the enhanced reasoning quality learned during RL training rather than from the output format itself.

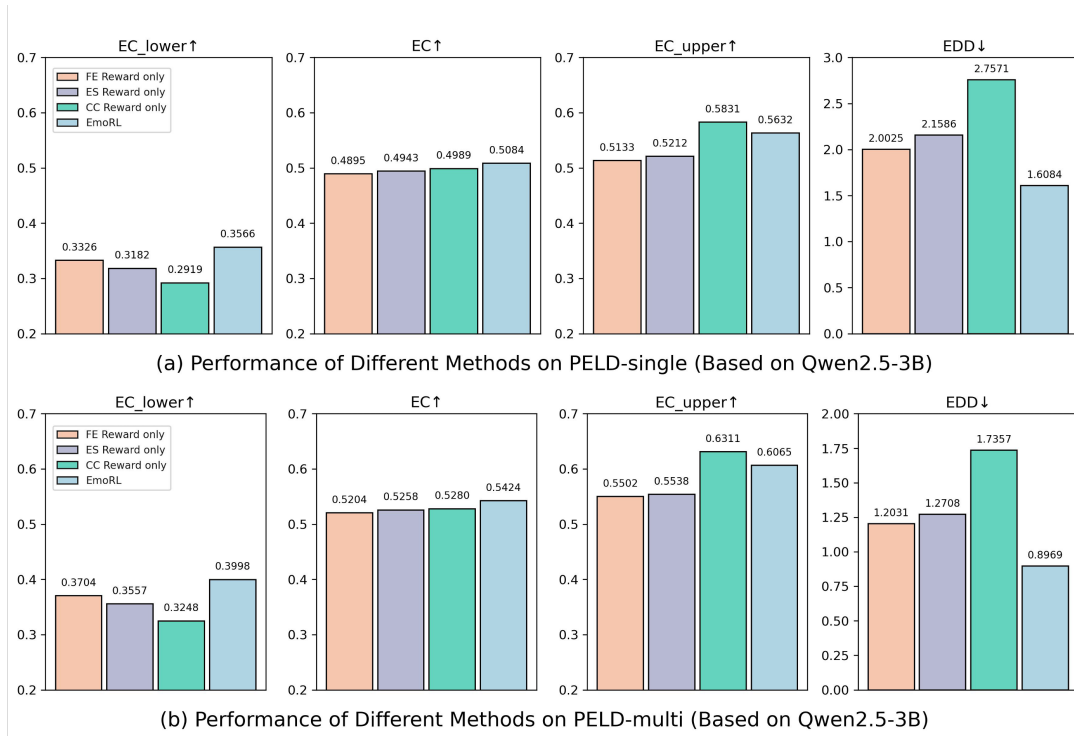


Figure 6: Performance results of RL with different reward on PELD-single and PELD-multi. ↑: Larger is better. ↓: Smaller is better.

Prompt Template for Emotion and Response (ER)

System Prompt:

{System_prompt_for_character R} + “When responding, use the JSON format, for example, {‘emotion’: ‘fear’, ‘content’: ‘No don’t, I beg of you!’}. ‘emotion’ represents the emotion used in the response, which has seven types: ‘anger’, ‘disgust’, ‘fear’, ‘joy’, ‘neutral’, ‘sadness’, ‘surprise’. ‘content’ represents the content of the reply.”

Assistant Prompt:

{‘emotion’: emotion_1, ‘content’: utterance_1}

User Prompt:

{‘emotion’: emotion_2, ‘content’: utterance_2}

...

Assistant Prompt:

{‘emotion’: emotion_n-2, ‘content’: utterance_n-2}

User Prompt:

{‘emotion’: emotion_n-1, ‘content’: utterance_n-1}

Assistant (waiting for responses from RPAs):

{RPA’s response}

Figure 7: Prompt Template for Emotion and Response (ER)

Main characters' Description

Chandler's Description:

Chandler Bing is known for his quick-witted and sarcastic humor, which defines much of his personality in the TV series *'Friends'*. He often uses humor as a shield to avoid dealing with his emotions and insecurities stemming from his parents' tumultuous divorce and his father's unconventional lifestyle. Despite his penchant for sarcasm, Chandler is also incredibly loyal and caring. He values his friendships deeply and is always ready to support his friends in times of need.

Joey's Description:

Joey Tribbiani from *'Friends'* is affable, simple-minded, and loyal. He's an aspiring actor, known for his love of food and optimistic outlook. Despite appearing naive, Joey is empathetic and caring. He often provides comic relief but also shows deep emotional depth. His charm and protective nature make him beloved among his friends, emphasizing his loyalty and affection.

Ross's Description:

Ross Geller from *'Friends'* is intellectual, sensitive, and socially awkward. He's a paleontologist known for his rationality and scientific mindset, contrasting with his friends' laid-back attitudes. He is loving and devoted, especially to his family and son, though he can be overly serious and jealous. His passion for dinosaurs and history, nerdy traits, and occasional anger add complexity to his character.

Monica's Description:

Monica Geller from *'Friends'* is competitive, highly organized, and obsessively clean. She is a professional chef, showcasing her passion and perfectionism. Monica often acts as the mother hen of her group, being nurturing yet sometimes overly controlling. Despite her strong-willed nature, she is loyal and humorous, especially about her compulsiveness. Her story includes personal growth, dealing with family and love issues. Monica's mix of strength, vulnerability, and maternal qualities make her a central and beloved character.

Phoebe's Description:

Phoebe Buffay from *'Friends'* is quirky, eccentric, and free-spirited. She's a masseuse and self-taught musician known for her humorous and bizarre songs. Phoebe's unconventional past, including living on the streets, shapes her unique worldview. She's empathetic, connected to spirituality, and refreshingly honest. Despite her oddities, Phoebe is fiercely loyal, compassionate, and protective of her friends. Her resilience and individuality make her a beloved and standout character in the series.

Rachel's Description:

Rachel Green from *'Friends'* starts as a spoiled, naive woman and evolves into an independent, successful fashion professional. Initially a waitress, she climbs the fashion industry ladder, showcasing her style and personal growth. Rachel is warm, affectionate, and romantic. She's also humorous and quirky, endearing her to her friends and the audience. Rachel values her friendships deeply, displaying loyalty and compassion. Her journey from dependency to independence, coupled with her loving nature, makes her a cherished character.

Figure 8: Descriptions of the main characters in *'Friends'*