EmoCharacter: Evaluating the Emotional Fidelity of Role-Playing Agents in Dialogues

Qiming Feng[1], Qiujie Xie[1], Xiaolong Wang[1], Qingqiu Li[1], Yuejie Zhang[1*], Rui Feng[1*], Tao Zhang[2*], Shang Gao[3]

[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University,
[2]School of Information Management and Engineering, Shanghai Key Laboratory of Financial Information Technology,
Shanghai University of Finance and Economics, [3]School of Information Technology, Deakin University
{qmfeng23, qjxie22, xiaolongwang24, qqli22}@m.fudan.edu.cn,
{yjzhang, fengrui}@fudan.edu.cn, taozhang@mail.shufe.edu.cn, shang.gao@deakin.edu.au

Abstract

Role-playing agents (RPAs) powered by large language models (LLMs) have been widely utilized in dialogue systems for their capability to deliver personalized interactions. Current evaluations of RPAs mainly focus on personality fidelity, tone imitation, and knowledge consistency, while overlooking emotional fidelity, a key factor that affects user experience. To this end, we propose a benchmark called EmoCharacter to assess emotional fidelity of RPAs in dialogues. EmoCharacter includes two benchmark datasets (single-turn and multi-turn dialogues), three evaluation settings, and six metrics to measure the emotional fidelity between RPAs and the characters they portray. Based on EmoCharacter, we conduct extensive evaluations on RPAs powered by seven widely used LLMs with representative role-playing methods. Our empirical findings reveal that: (1) Contrary to intuition, current role-playing methods often reduce the emotional fidelity of LLMs in dialogues; (2) Enhancing the general capabilities of LLMs does not necessarily improve the emotional fidelity of RPAs; (3) Fine-tuning or In-Context Learning based on real dialogue data can enhance emotional fidelity.

1 Introduction

The rise of large language models (LLMs) has revolutionized dialogue systems, enabling more dynamic and personalized interactions (Wang et al., 2023a; Yi et al., 2024). Among these advancements, role-playing agents (RPAs) have emerged, with agents assigned specific characters or personas. RPAs have been successfully deployed in various applications, ranging from entertainment to mental health support (Tseng et al., 2024; Park et al., 2023; Gan et al., 2023; Yang et al., 2024). Current evaluations of RPAs primarily focus on personality fidelity, tone imitation, and knowledge consistency (Wang et al., 2024b, 2023b; Shao et al., 2023), while overlooking a critical aspect of human-like interaction: emotional fidelity.

Emotional fidelity refers to the degree to which an RPA can simulate the emotional dynamics of the character it portrays. Human beings naturally display emotional shifts based on the context, tone, and flow of conversation (Chang et al., 2023; Chen et al., 2017). Psychological research indicates that personality is closely linked to emotions, with different characters exhibiting varying emotional dynamics in dialogues (Poria et al., 2018; Kokkonen and Pulkkinen, 1999). The ability to exhibit emotional variability and coherence throughout a dialogue is essential for achieving realistic and engaging role-playing experiences.

However, existing RPAs often fail to capture these nuances, offering static emotional responses or inaccurately portraying emotions, thereby breaking the illusion of authenticity. To address this gap, we propose EmoCharacter, a novel benchmark designed to assess the emotional fidelity of RPAs. EmoCharacter includes two benchmark datasets derived from classic TV dialogues and introduces six metrics across three evaluation settings to quantitatively measure how well RPAs maintain emotional fidelity with the characters they portray.

Through extensive experiments (§4) and analysis (§5) of RPAs powered by seven widely used LLMs with three different role-playing methods (None, RoleLLM and Chat-Haruhi), we pinpoint three key insights: (1) Contrary to intuition, role-playing methods often reduce the emotional fidelity of LLMs in dialogues; (2) Enhancing the general capabilities of LLMs does not necessarily improve the emotional fidelity in RPAs; (3) Supervised Fine-tuning (SFT) or In-Context Learning (ICL) using real dialogue data annotated with emotional states can enhance emotional fidelity.

Our contributions are as follows:

(1) To the best of our knowledge, we are the first to systematically study the emotional fidelity of

*Corresponding authors.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6218–6240

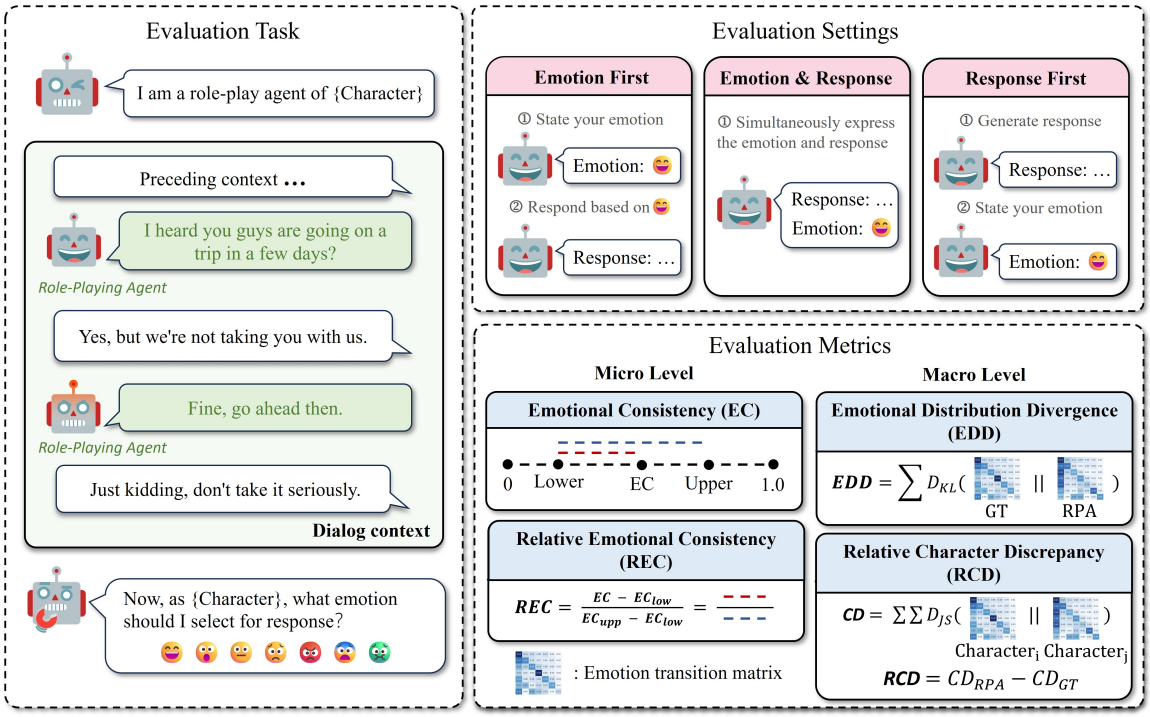

Figure 1: Framework of EmoCharacter. **Left**: format of evaluation tasks for assessing emotional fidelity in RPAs. **Upper right**: three evaluation settings. **Lower right**: metrics designed from both micro and macro perspectives.

RPAs in dialogue.

(2) We design six metrics to quantify emotional fidelity from both macro and micro perspectives and conduct extensive experiments on various LLMs with role-playing methods, drawing conclusions that benefit future research.

(3) We conduct an in-depth exploration of methods to enhance emotional fidelity, including SFT and ICL. Results indicate their efficacy in improving emotional fidelity.

2 Related Work

LLM-based Role-Playing. Role-playing agents (RPAs), typically powered by LLMs, are AI systems capable of simulating and interacting as their assigned roles (Chen et al., 2024). Current research mainly focuses on two directions: (1) Developing RPAs for specific roles, such as RoleLLM (Wang et al., 2023b) and Chat-Haruhi (Li et al., 2023), and (2) Training foundation models specialized in role-playing, such as Doubao-Character and character.ai (DoubaoTeam, 2024; character.ai, 2024). To enable RPAs to fully learn and utilize role data, various methods are employed to enhance the role-playing capabilities of LLMs, including setting the role description as a system prompt (Shao et al., 2023; Lu et al., 2024), training with role dialogue data (Shao et al., 2023), and retrieving relevant experiences from role databases (Wang et al., 2023b).

Evaluation of RPAs. Previous work on evaluating RPAs has mainly focused on two aspects: (1) Role fidelity, including personality fidelity (Wang et al., 2024b) and consistency in role knowledge, experience, and language patterns (Wang et al., 2023b; Shao et al., 2023), and (2) Natural abilities a role should possess, such as multi-turn dialogue capability (Duan et al., 2024), human-likeness (Xie et al., 2024), and attractiveness (Zhou et al., 2023). An ideal RPA should mimic the assigned character's tone, embody its personality, and respond with appropriate emotions to enhance immersion and appeal (Chen et al., 2024). However, current evaluations overlook emotional fidelity, a crucial aspect of emotional intelligence that significantly affects user experience, thereby motivating the work presented in this paper.

3 EmoCharacter

To evaluate the emotional fidelity between RPAs and their assigned characters, EmoCharacter incorporates three components, as shown in Figure 1: evaluation task and settings, single-turn and multi-turn dialogue datasets, and evaluation metrics.

3.1 Evaluation Task and Settings

Preliminaries. To assess the emotional fidelity of RPAs, we first need to obtain the responses and emotions generated by the RPA during dialogue. Formally, given a turn-taking dialogue between a target character R and a participant. We define the ground-truth dialogue as $C_n = \{U^1_{gt}, U^2_{gt}, \ldots, U^n_{gt}\}$ and the set of emotions associated with each utterance as $E_n = \{E^1_{gt}, E^2_{gt}, \ldots, E^{n-1}_{gt}, E^n_{gt}\}$. Both the first and last utterances, U^1_{gt} and U^n_{gt}, are from character R. The process of obtaining the responses and emotions generated by the RPA playing role R is defined as:

$$(U^n_R, E^n_R) = \text{RPA}(R, C_{n-1}, E_{n-1}) \quad (1)$$

We call each (R, C_{n-1}, E_{n-1}) triplet a test point.

Evaluation settings. Due to the autoregressive nature of LLMs, variations in the output format can result in significantly different generation outcomes in the subsequent text. To ensure that the obtained emotional response (U^n_R, E^n_R) of the RPA is stable and consistent with those in the RPA's regular dialogues, we design three evaluation settings to capture (U^n_R, E^n_R) from the conversation: (1) **Emotion First**: RPA first expresses its current emotion, followed by a response based on that emotion. (2) **Emotion and Response**: RPA simultaneously expresses its current emotion and response. (3) **Response First**: RPA first provides a response, then expresses the associated emotion.

Task definition. After obtaining the emotional response (U^n_R, E^n_R), we define the evaluation task as assessing the consistency between the RPA's dialogue $(U^n_R, E^n_R, C_{n-1}, E_{n-1})$ and ground truth dialogue (C_n, E_n) using the metrics from §3.3.

3.2 Single-turn and Multi-turn Dataset

3.2.1 PELD-single and PELD-multi

Our single-turn and multi-turn 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. Building on those, (Wen et al., 2021) introduced Personality EmotionLines Dataset (PELD). Each sample in PELD consists of a single-turn dialogue triplet $C_3 = \{U^1_{gt}, U^2_{gt}, U^3_{gt}\}$ with corresponding emotion labels $E_3 = \{E^1_{gt}, E^2_{gt}, E^3_{gt}\}$. We adopt PELD as our single-turn dialogue dataset, naming it PELD-single.

Since multi-turn dialogues are a more common application scenario for RPAs, we restructure the MELD dataset to extract all two-person multi-turn dialogues to address the lack of multi-turn data in PELD. This results in $C_n = \{U^1_{gt}, U^2_{gt}, \ldots, U^n_{gt}\}$, where $n > 3$. We refer to this restructured multi-turn dialogue dataset as PELD-multi. Figure 16 illustrates one example of a multi-turn dialogue.

Similar to the MELD and EmoryNLP datasets, both PELD-single and PELD-multi face the class imbalance issue. This imbalance reflects how humans typically express emotions in natural conversations, where the majority of dialogue tends to be emotionally neutral. Table 1 shows the specific statistics. Here, "sentiment" refers to a coarser-grained emotional state, while "emotion" is finer-grained (see Appendix A.1 for details).

Category	PELD-single	PELD-multi
# Basic Statistics		
Test Points	6510	1391
Avg. Uttr. Len.	9.32	7.67
Avg. Dialog Len.	3	6.92
# Emotion		
Joy	3533	1667
Surprise	1889	1341
Neutral	8701	4544
Anger	2340	968
Disgust	376	285
Fear	1346	266
Sadness	1345	601
# Sentiment		
Positive	5422	3008
Neutral	8701	4544
Negative	5407	2120

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

3.2.2 Emotion Transition in Dialogue

Characters with different personalities exhibit unique emotional changes during conversations, and these dynamic shifts cannot be captured through simple categorical statistics. Therefore, we further calculate the emotional transitions for each character in PELD-single and PELD-multi. Specifically, we define an emotional transition as the change from emotion E^i_{gt} associated with the previous utterance U^i_{gt} to emotion E^{i+2}_{gt} associated with the next utterance U^{i+2}_{gt} by character R in a dynamic conversation $C_n = \{U^1_{gt}, U^2_{gt}, \ldots, U^n_{gt}\}$, where $0 < i < n - 2$.

The emotion transition matrix is a probability transition matrix, which describes the probability

distribution of character R transitioning from the initial emotion E^i_{gt} to emotion E^{i+2}_{gt}. As shown in Figure 2, in the two transition matrices depicted, the darker areas are primarily concentrated in the first column and along the diagonal, indicating that most transitions occur towards a neutral emotion or remain the same. However, there are subtle differences in the matrices among different characters, especially in the transitions of negative emotions such as anger, sadness, fear, and disgust, where the differences between the two characters are more pronounced. This also indirectly suggests a correlation between personality differences and emotional changes. Complete illustration of the matrices are presented in Figures 5 and 6 in Appendix A.2.

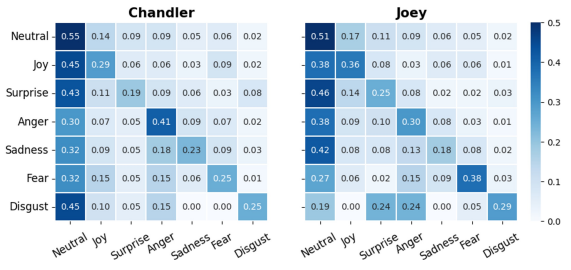

Figure 2: Emotion transition matrices for characters "Chandler" (sarcastic and insecure) and "Joey" (charming and naïve) from PELD-single .

3.3 Evaluation Metrics

To comprehensively assess the emotional fidelity of roles in dialogue, we design six metrics at both the micro and macro levels.

3.3.1 Micro Emotional Fidelity

At the micro level, we focus on the consistency between the emotions attached to each utterance by RPAs in the dialogue and those of the original characters. We introduce Emotional Consistency Score (EC), along with its lower bound (EC_{low}), upper bound (EC_{upp}), and Relative Emotional Consistency Score (REC), with Semantic Similarity (SS) added as a reference.

Emotional Consistency (EC). **EC** score quantifies the alignment between the emotions expressed by RPAs in each response and the emotions of the original characters. Traditional dialogue emotion recognition studies typically treat emotion alignment as a multi-class classification problem, using weighted F1 score as a metric for model performance (Zheng et al., 2023; Shi and Huang, 2023; Hu et al., 2023). However, the degree of difference between emotions varies. For instance, the emotional states of "surprise" are quite close to "joy" but significantly differ from "sadness". The F1 score does not effectively capture these differences between emotions. Therefore, we introduce the concept of emotional similarity (**ES**) based on the Valence-Arousal-Dominance (VAD) emotional state space (Russell and Mehrabian, 1977), where more similar emotions are closer in distance within the VAD space. We define the ES between different emotions E_i and E_j as:

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

where $\vec{E}_i$ represents the vector of E_i in the VAD space, and $||*||_2$ is the L2 norm of the vector.

When the RPA playing role R responds with emotion E^n_R based on context C_{n-1}, and E^n_R differs from the character R's true emotion E^n_{gt}, we calculate the emotionally consistent portion and inconsistent portion between E^n_{gt} and E^n_R as $ES(E^n_{gt}, E^n_R)$ and $1 - ES(E^n_{gt}, E^n_R)$, respectively. The final EC score is obtained as the weight F1 score sum of $ES(E^n_{gt}, E^n_R)$ and $1 - ES(E^n_{gt}, E^n_R)$. More details are provided in the Appendix B. The original weight F1 score is the loose lower bound of EC, denoted as EC_{low}.

Furthermore, we can calculate the EC score at a coarser granularity. The seven emotion categories (Table 1) in PLED-single and PLED-multi can be grouped into Positive, Neutral, and Negative sentiment categories. For the RPA playing role R, when its E^n_R and E^n_{gt} fall within the same sentiment category, they are considered the same emotion; otherwise, the calculation follows the same method as EC. We call this coarser-grained EC score the upper bound of EC, EC_{upp}.

In summary, EC_{low} is the traditional weight F1 score, EC is the weight F1 score adjusted by ES, and EC_{upp} is the EC for sentiment category classification. Their relationship is: $0 \leq EC_{low} < EC \leq EC_{upp} \leq 1$. EC_{upp}, as the coarsest emotional consistency metric, can be seen as the upper limit of RPAs' current emotional consistency capability. EC is the metric that best reflects user experience, while EC_{low} is the finest emotional consistency metric, reflecting the lower bound of the RPA's capability.

Relative Emotional Consistency (REC). When EC is closer to EC_{low}, it indicates that RPAs consistently select emotions that are less aligned with

the ground truth, even though they fall within the correct sentiment category. This suggests that while the RPAs have the right general emotional direction, they struggle to select the most granular emotion. It also implies that there is greater potential to improve EC through optimization techniques such as prompt engineering or in-context learning. Conversely, when EC approaches EC_{upp}, it shows that RPAs are better at selecting the most precise emotion within the correct sentiment category, indicating that the limiting factor for improving EC is the capability of the base model. To reflect how well RPAs are utilizing the capabilities of the underlying LLMs, we introduce **REC**, defined as:

$$\text{REC} = \frac{EC - EC_{low}}{EC_{upp} - EC_{low}} \tag{3}$$

As REC approaches 0, EC gets closer to its lower bound EC_{low}; conversely, as it approaches 1, EC nears its upper bound EC_{upp}.

Semantic Similarity (SS). We define semantic similarity (SS) between two utterances as the cosine similarity of their embedding vectors, which are generated by the text-embedding-3-small model (OpenAI, 2024).

3.3.2 Macro Emotional Fidelity

At the Macro Level, we focus on whether the emotional transition distribution of RPAs during dialogues aligns with that of the original characters, while also capturing the emotional differences between characters. To evaluate this, we introduce two metrics: the Emotional Distribution Divergence (EDD) score and the Relative Character Divergence (RCD) score.

Emotional Distribution Divergence (EDD). For a given character, the probabilities of different emotional transitions in dialogues vary, and the emotional transition probability matrix effectively quantifies this variation. Let M_{gt} be the emotional probability transition matrix of the target character R, and M_R be the emotional transition matrix generated by the RPA playing R in dialogues. We define the KL divergence from M_{gt} to M_R as the **EDD** score for the RPA playing R:

$$\text{EDD} = D_{KL}(M_{\text{gt}} \parallel M_R), \tag{4}$$

where $\parallel$ denotes the divergence between two probability distributions. A smaller EDD score indicates a higher consistency in emotional distribution. To reduce the impact of many zeros in the emotional probability transition matrix, we apply Laplace smoothing in the calculations.

Relative Character Divergence (RCD). As shown in Figure 2, the emotional transition probability matrices for the characters in PELD-single exhibit slight differences, reflecting the unique characteristics of each role. Ideally, RPAs should also demonstrate such differences, providing more personalized dialogues. Let M_{R_i} and M_{R_j} be the emotional transition matrices of two characters R_i and R_j, respectively. We define Character Divergence (CD) as the average Jeffreys divergence (Jeffreys, 1946) between their emotional transition matrices:

$$\text{CD} = \frac{\sum_i \sum_j J(M_{R_i} \parallel M_{R_j})}{N}, \tag{5}$$

where N is the number of characters in dataset.

Based on CD scores, the **RCD** score is further defined as the difference between the divergence scores of the RPA and the original role:

$$\text{RCD} = \text{CD}_{\text{RPA}} - \text{CD}_{\text{gt}} \tag{6}$$

When RCD is closer to 0, it indicates that the differences between roles played by RPAs align more closely with the original characters. If RCD > 0, it suggests that the RPAs overly emphasize the character differences. Conversely, if RCD < 0, it implies that the emotional distributions between the roles played by the RPAs are too similar, failing to capture their distinct personalities.

4 Experiments

4.1 Experimental Setup

Foundation LLMs for RPAs. We select seven commonly used LLMs from three categories to serve as the foundation LLMs for RPAs: (1) closed-source models from the GPT series, including GPT-4o, GPT-3.5, and GPT-4o-mini (Achiam et al., 2023), (2) Open-source models from the Llama series, such as Llama3-8b-instruct and Llama3-70b-instruct [1](Dubey et al., 2024), and (3) Models specifically trained for role-playing tasks, including Doubao-character (DoubaoTeam, 2024) and Westlake-7b-v2 (Senseable, 2023).

Prompting Strategies. For the role-playing methods, we use the system prompt-based approaches proposed in RoleLLM and Chat-Haruhi

[1]For convenience, we abbreviate Llama3-8b-instruct and Llama3-70b-instruct as Llama3-8b and Llama3-70b.

to portray the six main characters from *"Friends"*, and compare their performance with that of a blank LLM (where the model has no information about the character, referred to as "None").

Data. For each combination of LLMs and role-playing methods, we evaluate their performance across the three evaluation settings (§3.1) and report the results of the best-performing setting. To ensure consistency in test scale, we use 10% of PELD-single and 50% of PELD-multi as single-turn and multi-turn test sets, respectively, with the remaining data used for the fine-tuning (§5.1.1). See Appendix C for details.

4.2 Main Results

Table 2 presents the results of our evaluation of LLMs and RPAs using EmoCharacter. Based on these results, we observe that:

(1) In both single-turn or multi-turn dialogues, the EC scores of LLMs using role-playing methods are lower compared to blank LLMs. Additionally, the EDD scores tend to increase, with the exception of Doubao. For instance, GPT-4o model without role-playing (None) achieves an EC of 0.484, but this drops to 0.422 and 0.423 with RoleLLM and Chat-Haruhi role-playing prompts, respectively. This indicates that LLMs struggle with emotional expression during role-playing, resulting in reduced emotional fidelity.

(2) Regarding RCD, all LLMs show negative values when not engaged in role-playing and positive values when they are, indicating that RPAs indeed reflect the personality traits of characters. However, this also results in excessive differences between characters.

(3) Counterintuitively, enhancing the general capabilities of LLMs does not lead to an improvement in the emotional fidelity of RPAs. For example, among the foundation LLMs, GPT-3.5 shows higher EC than GPT-4o-mini, and LLaMA3-70b outperforms LLaMA3-8b in EC, yet both perform worse during role-playing than weaker LLMs.

(4) LLMs specifically optimized for role-playing tasks, such as Doubao and Westlake, demonstrate stronger emotional fidelity. For instance, Westlake achieves an EC score during role-playing that surpasses both GPT-4o and Doubao, attaining the highest EC_{upp}.

Table 3 presents the semantic similarity (**SS**) results of GPT-3.5 under different settings. It can be observed that while emotional fidelity (ES and EDD) decreases with the RoleLLM prompting strategy, the semantic similarity increases from 0.208 to 0.230. This suggests that RPAs indeed become more similar to the role in certain aspects, such as tone and catchphrases. However, when we specify the same emotion to match the role, as in the 'emo' setting in the table, the semantic similarity further improves to 0.289. This indicates that improving emotional fidelity can enhance the quality of responses.

4.3 Relative Emotional Consistency Analysis

Figure 3 illustrates the distribution of REC and EC for different LLMs under the None, RoleLLM, and Chat-Haruhi methods. Full results are available in Appendix C.3. REC indicates the position of EC within the interval (EC_{low}, EC_{upp}) and reflects the potential for LLMs to improve EC. We observe that Westlake maintains a high REC across all three methods, with scores of 70.7%, 72.6%, and 67.0%, respectively, achieving good EC even without a high EC_{upp}. In contrast, Doubao maintains a high EC_{upp} across all methods, but its EC is less impressive, resulting in relatively low REC scores of 52.0%, 55.9%, and 55.7%. This indicates that current role-playing methods have not yet fully utilized Doubao's potential, and Doubao has greater potential than Westlake to improve EC through prompting and other techniques. The limitation in further improving Westlake's EC may be due to the capabilities of its foundation model.

5 Analysis

5.1 How to Improve Emotional Fidelity?

5.1.1 Real Dialogue Data

Extensive experiments show that the emotional fidelity of LLMs decreases during role-playing (as shown in Table 2). To address this, we further explore whether using real dialogue data for SFT could improve emotional fidelity. Specifically, we construct three forms of fine-tuning datasets from 40% of PELD-single dataset according to the three task settings outlined in §3.1 (Emo-First, Emo & Resp, and Resp-First). We select Llama3-8b model for fine-tuning with LoRA (Hu et al., 2021), as it is a widely used general-purpose open-source model for role-playing, and has impressive capabilities despite having a relatively small number of parameters.

Table 4 presents the emotional fidelity metrics of the fine-tuned models. We observe that fine-

Model	Type	Single-turn Dialogue					Muti-turn Dialogue				
		Eomtion Consistency			EDD ↓	RCD ↔	Eomtion Consistency			EDD ↓	RCD ↔
		lower ↑	EC ↑	upper ↑			lower ↑	EC ↑	upper ↑		
GPT - series											
GPT-4o	None	**0.340**	**0.484**	**0.523**	**1.300**	**-0.260**	**0.343**	**0.467**	**0.553**	**1.554**	**-0.319**
	RoleLLM	0.312	0.422	0.489	1.667	2.801	0.300	0.419	0.500	1.813	2.617
	Chat-Haruhi	0.283	0.423	0.477	1.874	1.960	0.303	0.433	0.505	1.870	0.503
GPT-3.5	None	**0.318**	**0.439**	**0.506**	**1.007**	-1.905	**0.328**	**0.466**	**0.526**	**1.550**	**-0.778**
	RoleLLM	0.261	0.397	0.454	1.660	0.757	0.290	0.425	0.483	1.983	1.117
	Chat-Haruhi	0.275	0.399	0.473	1.874	**0.347**	0.310	0.426	0.486	1.710	1.352
GPT-4o-mini	None	**0.300**	**0.420**	**0.502**	**1.111**	-1.414	**0.321**	**0.436**	**0.502**	**1.364**	**-0.566**
	RoleLLM	0.264	0.404	0.473	1.760	**0.519**	0.285	0.414	0.475	1.838	1.790
	Chat-Haruhi	0.280	0.407	0.484	1.356	1.460	0.266	0.394	0.476	1.867	2.032
Llama - series											
Llama3-70b	None	0.287	**0.427**	0.494	1.367	-1.173	**0.309**	**0.437**	**0.500**	**1.822**	-1.265
	RoleLLM	0.169	0.356	0.434	2.910	**0.749**	0.243	0.402	0.472	2.399	1.360
	Chat-Haruhi	**0.317**	0.417	**0.516**	**1.244**	1.998	0.269	0.410	0.485	2.010	**0.507**
Llama-8b	None	0.260	**0.423**	**0.514**	1.742	-1.205	**0.303**	**0.436**	0.509	**1.845**	**-0.288**
	RoleLLM	0.218	0.408	0.489	2.766	**0.246**	0.224	0.382	0.462	2.511	4.037
	Chat-Haruhi	**0.307**	0.414	0.501	**1.260**	2.001	0.256	0.430	**0.523**	1.942	3.322
Role Play LLMs											
Doubao-character	None	**0.313**	**0.430**	**0.538**	1.410	-0.886	**0.288**	**0.423**	**0.535**	2.420	**0.206**
	RoleLLM	0.290	0.409	0.503	1.185	1.305	0.246	0.384	0.490	2.187	1.094
	Chat-Haruhi	0.288	0.406	0.500	**1.159**	**0.646**	0.257	0.391	0.497	**2.009**	0.524
Westlake-7b	None	**0.326**	**0.471**	**0.531**	**1.152**	-2.440					
	RoleLLM	0.303	0.433	0.482	1.346	0.267			*		
	Chat-Haruhi	0.303	0.429	0.491	1.348	**-0.263**					

Table 2: Performance results of base LLMs with role-playing methods on PELD-single and PELD-multi. Best results are in bold. ↔: The closer to zero, the better. *: Results for Westlake are not presented for fairness, as it frequently deviates from the setup (e.g., providing irrelevant responses). See Appendix C.3 for complete results.

RPA Method	EC ↑	EDD ↓	SS ↑
GPT-3.5-None	0.439	1.007	0.208 ,
RoleLLM (GPT-3.5)	0.399	1.660	0.230
RoleLLM (GPT-3.5, emo)	**1.000**	**0.000**	**0.289**

Table 3: Results of EC, EDD, and SS for LLMs in single-turn dialogues. "emo" indicates that the LLM is provided with ground truth emotion prior to responding.

Model	SFT Dataset	Eomtion Consistency			EDD ↓	RCD ↔
		lower ↑	EC ↑	upper ↑		
GPT-4o	/	0.312	0.422	0.489	**1.667**	2.801
Llama3-8b	/	0.218	0.408	0.489	2.766	**0.246**
Llama3-8b -sft	Emo - First	0.284	0.454	0.489	1.973	-1.678
	Emo & Resp	**0.340**	**0.499**	**0.551**	1.680	-0.556
	Resp - First	0.311	0.468	0.506	1.695	-0.717

Table 4: Results of fine-tuning Llama3-8b using three different datasets.

tuning LLMs with real dialogue data can significantly enhance their EC, EC_{low}, EC_{upp}, and EDD during role-playing, even surpassing powerful general models like GPT-4o. However, fine-tuning reduces the role differentiation of LLMs, with RCD becoming negative regardless of the fine-tuning approach. See Appendix C.4 for fine-tuning details.

5.1.2 In-Context Learning (ICL)

Existing RPA work often enhances role-playing abilities by retrieving a character's historical experiences (such as related dialogues from a character database) during interactions, leveraging the ICL capability of LLMs. To further explore the impact of different ICL settings on the emotional fidelity of RPAs, we select 192 test points from PELD-multi and define n-shot as adding n rounds of real dialogues before the test point as a reference. Table 5 shows the performance of RPAs using the RoleLLM and Chat-Haruhi prompting strategies in terms of EC, EC_{low}, EC_{upp}, and REC under different shot settings. We observe that as the number of shots increases, the model's EC shows an upward trend, but the improvement is limited.

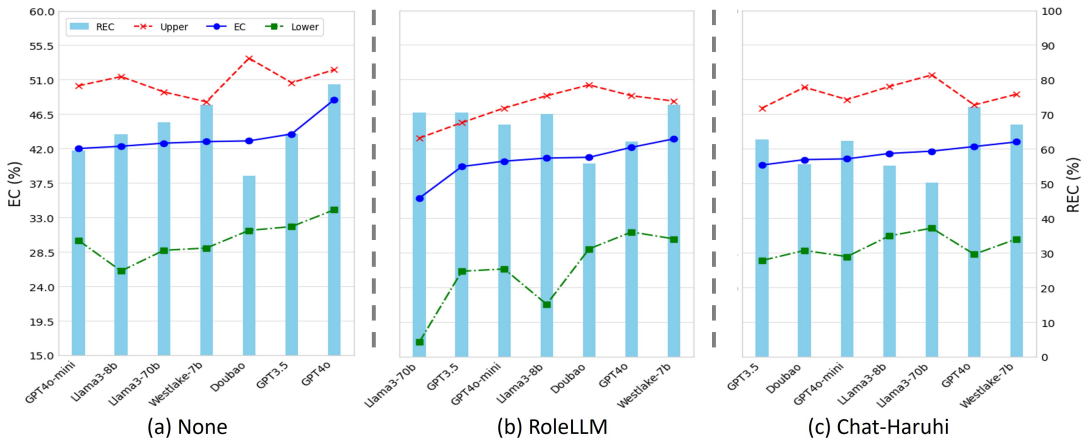

Figure 3: Relative emotional consistency results for different role-playing methods on PELD-single (ordered by EC in ascending order). Besides REC bar charts, colored lines show EC and its upper and lower bounds for comparison.

Shot	lower ↑	EC ↑	upper ↑
RoleLLM (GPT-3.5 base)			
0-shot	0.276	0.410	0.488
1-shot	**0.336**	0.441	0.514
2-shot	0.309	0.436	**0.544**
3-shot	0.330	**0.445**	0.525
Chat-Haruhi (GPT-3.5 base)			
0-shot	0.234	0.375	0.480
1-shot	0.269	0.389	0.480
2-shot	0.269	0.392	0.475
3-shot	**0.307**	**0.418**	**0.495**

Table 5: Emotional Consistency of GPT-3.5 with RoleLLM and Chat-Haruhi under 0-3 ICL shot settings.

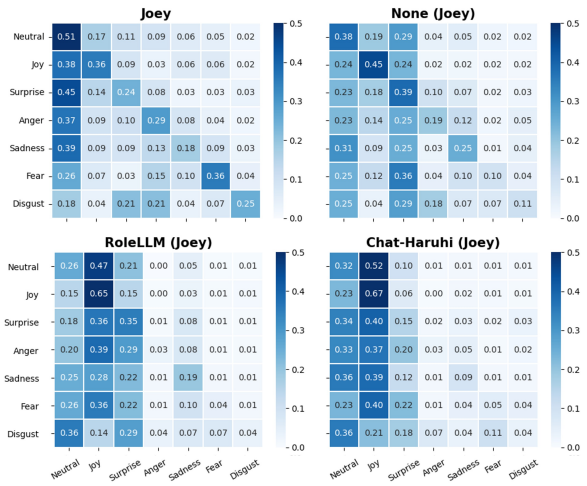

Figure 4: Real emotion transformation matrix of "Joey" in PELD-single and matrices generated using RoleLLM and Chat-Haruhi. "None" indicates use of blank LLMs.

5.2 Case Study: Why Emotional Infidelity?

In extensive experiments, we observe that even the most advanced LLMs struggle to maintain high emotional fidelity during role-playing dialogues. The emotional fidelity of LLMs is often lower than when they are not engaging in role-playing. To explore the reasons behind this decline and the emotional discrepancies between RPAs and real characters, we conduct a case study. We test the GPT-3.5 model using RoleLLM and Chat-Haruhi on the complete PELD-single dataset. Figure 4 shows the original emotion transition matrix for character "Joey" and the matrices for the LLM itself (None) and when it plays the character of "Joey" using RoleLLM and Chat-Haruhi.

As seen in Figure 4, the first three columns of the emotional transition matrices for RoleLLM and Chat-Haruhi are notably dark, indicating that RPAs tend to use neutral and positive emotions like 'Neutral', 'Joy', and 'Surprise' in their responses, while rarely employing negative emotions. This significantly diverges from the character's original emotional state, which may result from emotional bias introduced during the RLHF (Reinforcement Learning from Human Feedback) phase of training LLMs (e.g., certain emotional biases are already evident in the emotion transition matrix for None.) (Bai et al., 2022; Ouyang et al., 2022; Cao et al., 2024), or from an over-adherence to the system prompt (as character descriptions often emphasize positive traits) (Wang et al., 2024a; Lou et al., 2024). See more case studies in Appendix C.5.

6 Conclusion

We proposed EmoCharacter to evaluate the emotional fidelity of RPAs from both macro and micro perspectives. With extensive evaluations of multiple LLMs and role-playing methods, we pinpointed that even state-of-the-art LLMs like GPT-4o experience a decline in emotional fidelity during role-playing dialogues. Our case studies revealed that LLMs often use overly distinct emotions to

express character traits, leading to low emotional fidelity. Additionally, we found that fine-tuning with real dialogue data and using in-context learning both benefit the RPAs' emotional fidelity. We hope EmoCharacter will facilitate research on building RPAs with greater emotional intelligence.

7 Limitations

In this study, we conducted a comprehensive evaluation of the emotional fidelity of RPAs, but some limitations remain. Firstly, to obtain emotional labels for each response in RPA dialogues, we established three test settings to ensure the validity of our conclusions as much as possible, though biases may still exist. Secondly, due to computational constraints, we only fine-tuned Llama3-8b with LoRA in §5.1.1. Future efforts will aim to extend these results to larger LLMs, such as Llama3-70b. Lastly, this study focuses solely on the text modality, whereas in real emotional interactions, information such as voice, gestures, and facial expressions is equally crucial. Therefore, future research will expand to multimodal dialogue.

8 Ethical Considerations

In this study, we used dialogue datasets sourced from publicly available and widely researched resources, and the LLMs employed were accessed via publicly available APIs or parameters. Therefore, we do not anticipate any ethical issues arising from the research. Additionally, evaluating the emotional fidelity of RPAs can reveal their emotional state distribution and, to some extent, predict the probability of generating toxic content. If RPAs exhibit a significant amount of negative emotions, caution should be exercised, and measures should be taken to mitigate associated risks. Lastly, we confirm that all authors are aware of and adhere to the ACL ethics guidelines.

Acknowledgments

This work is supported by Natural Science Foundation of China (No. 62172101), the Science and Technology Commission of Shanghai Municipality (No. 23511100602), and the Fundamental Research Funds for the Central Universities of China (No. 2023110139).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Guolong Liu, Gaoqi Liang, Junhua Zhao, and Yun Li. 2024. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *ArXiv*, abs/2404.00282.

Che-Jui Chang, Samuel S. Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. 2023. The importance of multimodal emotioth conditioning and affect consistency for embodied conversational agents. *Proceedings of the 28th International Conference on Intelligent User Interfaces*.

character.ai. 2024. Personalized ai for every moment of your day. https://character.ai/about. Accessed: 2024-10-10.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ArXiv*, abs/1711.01731.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

DoubaoTeam. 2024. Doubao-pro. https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-pro-4k. Accessed: 2024-10-10.

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, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Paul Ekman and Richard J. Davidson. 1994. The nature of emotion: Fundamental questions.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. *2023 IEEE International Conference on Big Data (BigData)*, pages 4776–4785.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Harold Jeffreys. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Marja Kokkonen and Lea Pulkkinen. 1999. Emotion regulation strategies in relation to personality characteristics indicating low and high self-control of emotions. *Personality and Individual Differences*, 27:913–932.

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.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

OpenAI. 2024. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: 2024-10-10.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

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.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Senseable. 2023. Westlake-7b-v2: Role-play & text generation specialist model. https://huggingface.co/senseable/WestLake-7B-v2. Accessed: 2024-10-10.

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.

Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766, Toronto, Canada. Association for Computational Linguistics.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *ArXiv*, abs/2406.01171.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jing Zhou, Yufei Wang, and Kam-Fai Wong. 2023a. A survey of the evolution of language model-based dialogue systems. *ArXiv*, abs/2311.16789.

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. 2024a. A survey on data selection for llm instruction tuning. *ArXiv*, abs/2402.05123.

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. 2024b. 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, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

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, Online. Association for Computational Linguistics.

Qiuejie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang. 2024. Human simulacra: Benchmarking the personification of large language models. *arXiv preprint arXiv:2402.18180*.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *ArXiv*, abs/2406.16442.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *ArXiv*, abs/2402.18013.

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

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459, Toronto, Canada. Association for Computational Linguistics.

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

A More Details on Datasets

A.1 Categories of Emotion and VAD Space

In this paper, we classify emotions in dialogue into six basic categories: *anger, disgust, fear, joy, sadness, surprise*, along with the intermediate emotion, *neutral* (Ekman and Davidson, 1994). Furthermore, we group these emotions into Positive, Negative, and Neutral categories. Russell and Mehrabian (1977) projected these emotions onto the Valence-Arousal-Dominance (VAD) space, quantifying the intensity of emotions across three dimensions. Details are presented in Table 6.

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 6: Emotions in VAD space

A.2 Emotion probability transition matrices in PELD-single and PELD-multi

Figures 5 and Figures 6 show the emotion transition matrices for the six main characters in the PLED-single and PLED-multi datasets, respectively.

B More Details on Evaluation Metrics

Lower bound of EC (EC_{low}). In §3.3.1 we use the weighted F1 score as EC_{low}. Specifically, for each emotion category, we calculate its true positives (**TP**), false positives (**FP**), and false negatives (**FN**). Here, **TP** refers to the number of samples where the actual value is positive and is also predicted as positive by the classifier. **FP** refers to the number of samples where the actual value is negative but is predicted as positive by the classifier. **FN** refers to the number of samples where the actual value is positive but is predicted as negative by the classifier. Then, we calculate $precision$ and $recall$ for each emotion category as follows:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{9}$$

Let the number of samples in the seven different emotion categories be $n_1, n_2, \ldots, n_7$, and their respective F1 scores be $F1_1, F1_2, \ldots, F1_7$. We derive EC_{low} by calculating the weighted F1 score

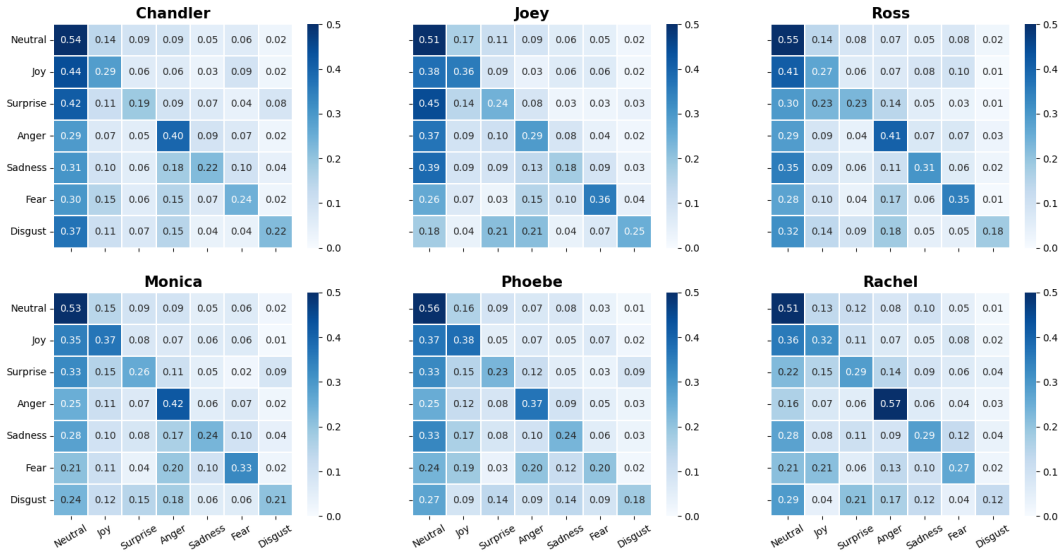

Figure 5: Emotion transition matrices of the main characters in PELD-single.

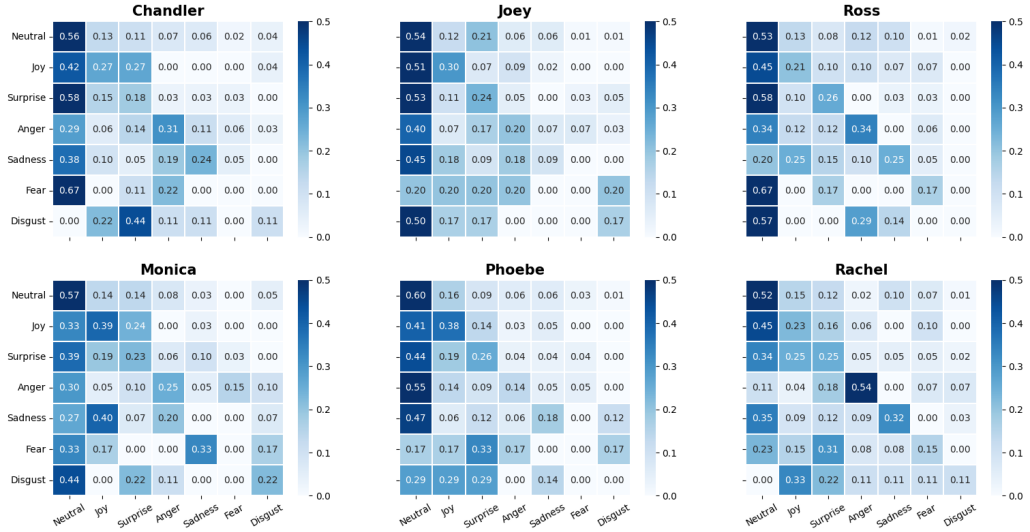

Figure 6: Emotion transition matrices for the main characters in PELD-multi.

as follows:

$$EC_{\text{low}} = \text{Weighted F1} = \frac{\sum_{i=1}^{7} n_i \times F1_i}{\sum_{i=1}^{7} n_i} \quad (10)$$

EC. In our calculation of EC_{low}, we review the method for computing the weighted F1 score. It is evident that EC_{low} serves as a stringent metric due to its strict evaluation criteria. For example, when the target emotion is "joy", any prediction of a different emotion is automatically classified as a FP, leading to an increment in the FP count while the TP count remains unchanged. Only when the predicted emotion matches the target "joy" does the TP count increase by one. This stringent criterion limits the metric's ability to account for potential correlations among emotions. As illustrated in Figure 7, the similarity between "joy" and "surprise" is evidently higher than that between "joy" and "sadness", yet such nuances are not captured by the current approach.

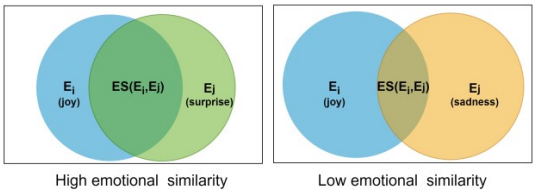

Figure 7: Emotional similarity between different emotions. The left diagram shows a high emotional similarity between "joy" (E_i) and "surprise" (E_j), indicated by a larger overlap ($ES(E_i, E_j)$). The right diagram illustrates a lower similarity between "joy" and "sadness", with a smaller overlap, highlighting varying degrees of emotional correlation.

To address this, we introduce ES to quantify

the degree of similarity between different emotions and adjust TP, FP, and FN accordingly. For instance, if "joy" is predicted as "surprise" in a sample, the TP for "joy" category is updated to $TP + ES(joy, surprise)$, FP is updated to $FP + (1 - ES(joy, surprise))$, and the FN for "surprise" category is updated to $FN + (1 - ES(joy, surprise))$. We denote these adjusted values as TP', FP', and FN'. Based on these adjusted values, we calculate the ES-adjusted F1 score, F1$'$, using Eq.9. Finally, the formula for calculating EC is:

$$EC = \frac{\sum_{i=1}^{7} n_i \times F1_i'}{\sum_{i=1}^{7} n_i} \tag{11}$$

From Eq. 2, it follows that $0 < ES < 1$, leading to the inequalities:

$$\begin{cases} TP' > TP \\ FP' < FP \\ FN' < FN \end{cases} \Rightarrow \text{F1}' > \text{F1} \Rightarrow ES > ES_{\text{low}}. \tag{12}$$

Therefore, we refer to EC_{low} as a loose lower bound of EC. Since EC considers the similarity between emotions, it better reflects the true emotional consistency of RPAs in dialogue, more accurately representing the user's real experience.

Upper bound of EC (EC_{upp}). As shown in Figure 6, the seven emotions can be grouped into three coarse-grained sentiments: Positive, Neutral, and Negative. Given the ground truth emotion E_{gt} and the emotion generated by the RPA, E_{RPA}, the evaluation approach depends on whether E_{gt} and E_{RPA} fall within the same sentiment category. If they belong to different sentiment groups, the calculation follows the original EC metric. Conversely, if E_{gt} and E_{RPA} are within the same sentiment group, they are treated as the same emotion. For instance, when E_{gt} is "joy" and E_{RPA} is "surprise", the true positive (TP) count for the "joy" category is incremented by one. However, if E_{RPA} is "sadness", the calculation reverts to the original EC metric. This modified metric, EC_{upp}, can be interpreted as a sentiment-consistency (coarse-grained) adjusted version of the original EC, representing the upper bound of the RPA's performance.

EDD and RCD In §3.3.2, we use KL divergence between probability matrices to calculate EDD. The KL divergence between probability matrices M_1 and M_2 is defined as:

$$D_{KL}(M_1 \parallel M_2) = \sum_i \sum_j M_1^{ij} \log \left(\frac{M_1^{ij}}{M_2^{ij}} \right), \tag{13}$$

Since KL divergence is asymmetric, we use Jeffreys divergence to calculate RCD:

$$J(M_1 \parallel M_2) = \frac{1}{2} D_{KL}(M_1 || M_2) + \frac{1}{2} D_{KL}(M_2 || M_1) \tag{14}$$

Validity of the metric To further validate the effectiveness of our evaluation framework, we conducted additional experiments to verify the similarity between experimental conclusions and human preferences. We enlisted three graduate student volunteers to score the model's responses by comparing them with the ground truth (i.e., the original character's responses). The volunteers were asked to determine which response's emotional expression was more aligned with the emotions the target character might likely exhibit (scoring from 0 to 1; if one response received a score of x, the other would receive a score of 1-x). Table 7 presents the results of the human preferences.

Method	EC_low	EC	EC_upp	HF
None	**0.318**	**0.439**	**0.506**	**0.382**
RoleLLM	0.261	0.397	0.454	0.264
Chat-haruhi	0.275	0.399	0.473	0.270

Table 7: Human feedback (HF) on responses from different methods. (based on GPT-3.5)

Application strategy of metrics or evaluation frameworks. Since our metrics rely on manually annotated emotion labels in the dialogue dataset as ground truth, there are instances where this ground truth might be missing. To address this, we propose two strategies to enhance the applicability of our evaluation framework: (1) using a dialogue emotion recognition model to generate pseudo-labels for dialogues. (2) Employing a character dataset with ground truth as a validation set to optimize the emotional fidelity of role-playing methods, and then using the optimized methods to portray the target character.

C More Details and Results from Experiments

C.1 Versions and System Prompts of LLMs

Table 8 shows the versions of all LLMs used in the experimental section. Figure 8 shows the prompts used in the RoleLLM and Chat-Haruhi prompting strategies. Figures 9 and 10 display the descriptive information for different characters.

Model	Version
GPT - series	
GPT-4o	gpt-4o-2024-05-13
GPT-3.5	gpt-3.5-turbo-0125
GPT-4o-mini	gpt-4o-mini-2024-07-18
Llama - series	
Llama3-8b	Llama3-8b-instruct
Llama3-70b	Llama3-70b-instruct
Role Play LLMs	
Westlake-7b	WestLake-7B-v2
Doubao-character	Doubao-pro-4k-character-240728

Table 8: Versions of LLMs.

C.2 More Details for Evaluation Settings

Figures 11, 12, and 13 show the prompts used in different evaluation settings. In fact, to refine the experiment further, the Emotion and Response (ER) in the evaluation settings can be be further broken down into Simultaneous-Emotion First (SEF) and Simultaneous-Response First (SRF). However, since the results obtained from these two setups are quite similar in most scenarios, we use SEF as the ER result.

C.3 Complete Experimental Results.

Tables 9, 10, and 11 present the full test results of different series of LLMs on PELD-single, while Tables 12, 13, and 14 show the complete test results on PELD-multi. Here, **EF** stands for "Emotion First", **ER** for "Emotion and Respond", and **RF** for "Respond First". We highlight the relatively better results from these three test settings in the main results. The error rate indicates the proportion of instances where the model's response contains emotions outside our predefined range or irrelevant content. We do not consider results with an error rate exceeding 10%. The REC results for multi-turn dialogues are shown in Figure 14.

C.4 More Details on the Supervised Fine-Tuning.

We use LoRA to fine-tune Llama3-8b-instruct with the following parameter settings: *epoch*: 3; *att_dropout*: 0.1; *learning_rate*: 0.00001; *lora_alpha*: 64; *lora_dropout*: 0.1; *lora_rank*: 64; *warmup_step_rate*: 0.05.

C.5 Additional Case Studies.

Figure 15 presents additional case studies.

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.

Prompt for Chat-Haruhi

I want you to act like {role_name} from *'Friends'*. I want you to respond and answer like {role_name} using the tone, manner, and vocabulary {role_name}. Your description is: {role_description}.

Figure 8: Prompts used in RoleLLM and Chat-Haruhi prompting strategies.

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.

Figure 9: Descriptions of the three main characters (Chandler, Joey, and Ross) in *'Friends'*

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 10: Descriptions of the other three main characters (Monica, Phoebe, and Rachel) in *'Friends'*

Prompt Template for Emotion First (EF)

System Prompt:
{System_prompt_for_character R}
Assistant Prompt:
{utterance_1}
User Prompt:
{utterance_2}
...
Assistant Prompt:
{utterance_n-2}
User Prompt:
{utterance_n-1}
User Prompt:
Please state the emotion you are experiencing after the previous exchange, selecting one of the following seven emotions: {'anger', 'disgust', 'fear', 'joy', 'neutral', 'sadness', 'surprise'}. !!Note: do not respond with any emotion categories other than these seven, simply answer with the word for the emotion you choose.
Assistant Prompt (waiting for responses from RPAs):
{RPA's response_1}
User Prompt:
Please respond to the previous exchange using the emotion of {RPA's response_1}.
Assistant (waiting for responses from RPAs):
{RPA's response_2}

Figure 11: Prompt Template for Emotion First (EF)

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 12: Prompt Template for Emotion and Response (ER)

Prompt Template for Response First (RF)

System Prompt:
{System_prompt_for_character R}
Assistant Prompt:
{utterance_1}
User Prompt:
{utterance_2}
...
Assistant Prompt:
{utterance_n-2}
User Prompt:
{utterance_n-1}
Assistant Prompt (waiting for responses from RPAs):
{RPA's response_1}
User Prompt:
Please state the emotion you attached to your response to the previous sentence, selecting one of the following seven emotions:{'anger', 'disgust', 'fear', 'joy', 'neutral', 'sadness', 'surprise'}. !!Note: do not respond with any emotion categories other than these seven, simply answer with the word for the emotion you choose."
Assistant (waiting for responses from RPAs):
{RPA's response_2}

Figure 13: Prompt Template for Response First (RF)

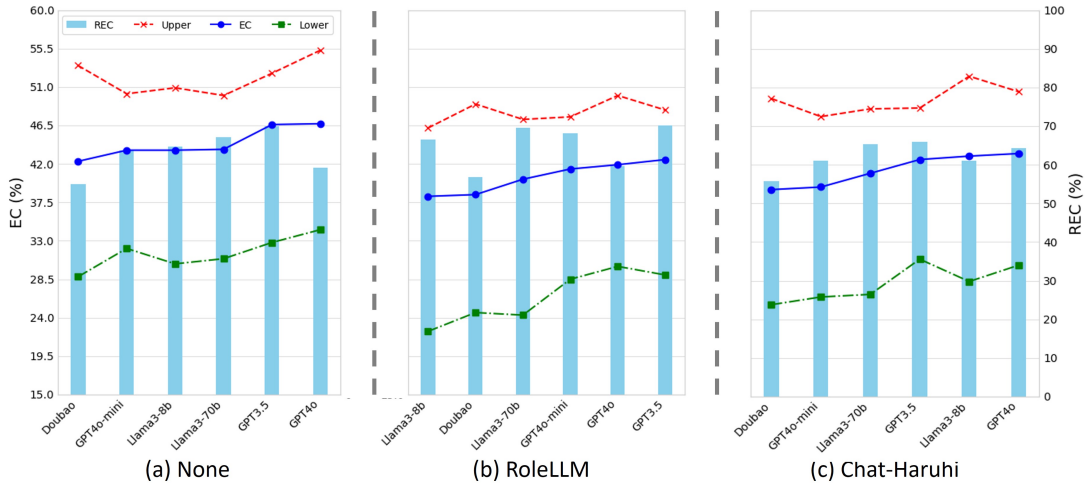

Figure 14: Relative emotional consistency results for different role-playing methods on PELD-multi (ordered by EC in ascending order). Besides REC bar charts, colored lines show EC and its upper and lower bounds for comparison.

Model	Type		Eomtion Consistency				EDD	RCD	Error rate
			lower	EC	upper	REC			
GPT4o	None	EF	0.336	0.479	0.521	0.773	**1.286**	-1.183	**0.005**
		ER	**0.340**	**0.484**	**0.523**	**0.787**	1.300	**-0.260**	0.027
		RF	0.317	0.454	0.505	0.729	1.352	-2.605	0.008
	RoleLLM	EF	**0.312**	0.422	0.489	0.621	**1.667**	2.801	**0.008**
		ER	0.306	**0.424**	**0.495**	0.624	1.672	4.582	0.018
		RF	0.304	0.418	0.471	**0.683**	1.682	**1.422**	0.016
	Chat-Haruhi	EF	**0.292**	0.419	0.475	0.694	**1.418**	1.691	**0.002**
		ER	0.283	**0.423**	**0.477**	**0.722**	1.874	1.960	0.029
		RF	0.257	0.390	0.459	0.658	1.773	**1.267**	0.005
GPT3.5	None	EF	**0.318**	**0.439**	0.506	0.644	**1.007**	-1.905	**0.000**
		ER	0.317	0.432	**0.509**	0.599	1.047	**-0.621**	0.003
		RF	0.286	0.426	0.474	**0.745**	1.517	-3.379	0.005
	RoleLLM	EF	**0.261**	**0.397**	**0.454**	**0.705**	**1.660**	**0.757**	0.008
		ER	0.256	0.373	0.442	0.629	1.812	4.283	**0.002**
		RF	0.250	0.376	0.449	0.633	1.858	3.670	0.010
	Chat-Haruhi	EF	**0.275**	**0.399**	**0.473**	0.626	**1.360**	**0.347**	0.006
		ER	0.256	0.371	0.442	0.618	1.930	4.263	**0.003**
		RF	0.254	0.383	0.455	**0.642**	1.731	1.860	0.005
GPT4o-mini	None	EF	**0.300**	**0.420**	0.502	0.594	**1.111**	-1.414	**0.009**
		ER	0.281	0.398	**0.509**	0.513	1.452	**-0.348**	0.012
		RF	0.277	0.414	0.472	**0.703**	1.487	-3.054	0.067
	RoleLLM	EF	**0.264**	**0.404**	**0.473**	**0.670**	**1.760**	**0.519**	0.017
		ER	0.169	0.315	0.416	0.591	2.544	3.862	0.015
		RF	0.226	0.361	0.433	0.652	1.978	4.363	**0.012**
	Chat-Haruhi	EF	**0.280**	**0.407**	**0.484**	**0.623**	**1.356**	**1.460**	**0.000**
		ER	0.185	0.322	0.410	0.609	2.529	1.890	0.029
		RF	0.240	0.371	0.459	0.598	1.739	1.993	0.015

Table 9: Evaluation results of GPT series models on PELD-single. **EF** stands for "Emotion First", **ER** for "Emotion and Respond", and **RF** for "Respond First". **Error rate** indicates the proportion of instances where the model's response contains emotions outside predefined range or irrelevant content.

Model	Type		Eomtion Consistency				EDD	RCD	Error rate
			lower	EC	upper	REC			
Llama3-70b	None	EF	0.287	**0.427**	**0.494**	0.676	**1.367**	**-1.173**	0.033
		ER	0.214	0.358	0.461	0.583	1.724	-2.067	0.178
		RF	**0.296**	**0.427**	0.472	**0.744**	1.418	-2.245	**0.005**
	RoleLLM	EF	0.169	**0.356**	**0.434**	**0.706**	2.910	**0.749**	0.012
		ER	**0.205**	0.350	0.420	0.674	**2.281**	3.479	0.315
		RF	0.174	0.320	0.396	0.658	2.858	3.753	**0.011**
	Chat-Haruhi	EF	**0.317**	**0.417**	**0.516**	0.503	**1.244**	**1.998**	**0.009**
		ER	0.230	0.402	0.512	**0.610**	2.500	7.904	0.263
		RF	0.262	0.396	0.486	0.598	1.864	6.920	0.015
Llama3-8b	None	EF	0.260	**0.423**	**0.514**	**0.642**	1.742	-1.205	0.022
		ER	0.235	0.396	0.491	0.629	1.780	-2.603	0.336
		RF	**0.303**	0.414	0.509	0.539	**1.474**	**-0.896**	**0.018**
	RoleLLM	EF	0.218	0.408	0.489	**0.701**	2.766	**0.246**	0.075
		ER	**0.249**	**0.416**	**0.513**	0.633	**2.085**	4.761	0.105
		RF	0.219	0.371	0.457	0.639	2.225	4.804	**0.024**
	Chat-Haruhi	EF	**0.307**	**0.414**	0.501	0.552	**1.260**	**2.001**	**0.006**
		ER	0.230	0.403	**0.511**	**0.616**	2.475	7.036	0.261
		RF	0.265	0.401	0.495	0.591	1.815	6.862	0.017

Table 10: Evaluation results of Llama series models on PELD-single.

Model	Type		Eomtion Consistency				EDD	RCD	Error rate
			lower	EC	upper	REC			
Doubao-character	None	EF	0.278	0.418	0.526	0.565	**1.346**	-1.024	0.006
		ER	**0.313**	**0.430**	**0.538**	0.520	1.410	-0.886	0.017
		RF	0.286	0.424	0.508	**0.622**	1.458	**-0.331**	**0.005**
	RoleLLM	EF	0.257	0.402	**0.517**	0.558	1.606	**0.098**	**0.006**
		ER	0.211	0.345	0.470	0.517	2.161	0.639	0.020
		RF	**0.290**	**0.409**	0.503	**0.559**	**1.185**	1.305	0.017
	Chat-Haruhi	EF	**0.288**	**0.406**	**0.500**	**0.557**	**1.159**	0.646	**0.000**
		ER	0.200	0.334	0.456	0.523	2.221	**0.511**	0.029
		RF	0.256	0.381	0.486	0.543	1.656	1.413	0.023
Westlake-7b	None	EF	0.290	0.429	0.481	0.728	1.339	**-1.492**	**0.049**
		ER	**0.326**	**0.471**	**0.531**	0.707	**1.152**	-2.440	0.457
		RF	0.272	0.411	0.453	**0.768**	1.517	-2.444	0.083
	RoleLLM	EF	**0.303**	0.433	0.482	0.726	**1.346**	**0.267**	**0.065**
		ER	0.278	**0.438**	**0.491**	**0.751**	1.602	-2.148	0.598
		RF	0.210	0.332	0.389	0.682	2.288	-0.384	0.066
	Chat-Haruhi	EF	**0.303**	**0.429**	**0.491**	0.670	1.348	**-0.263**	**0.021**
		ER	0.282	0.423	0.476	**0.727**	**1.314**	-1.107	0.387
		RF	0.188	0.314	0.372	0.685	2.259	-0.550	0.037

Table 11: Evaluation results of role-playing LLMs on PELD-single. **EF** stands for "Emotion First", **ER** for "Emotion and Respond", and **RF** for "Respond First". **Error rate** indicates the proportion of instances where model's response contains emotions outside predefined range or irrelevant content.

Model	Type		Eomtion Consistency				EDD	RCD	Error rate
			lower	EC	upper	REC			
GPT4o	None	EF	0.338	0.456	0.528	**0.621**	**1.423**	**-0.035**	**0.006**
		ER	0.308	0.432	0.531	0.556	1.929	0.903	0.026
		RF	**0.343**	**0.467**	**0.553**	0.590	1.554	-0.319	**0.006**
	RoleLLM	EF	0.294	**0.419**	0.493	**0.628**	**1.775**	**1.066**	**0.001**
		ER	0.284	0.413	**0.511**	0.568	2.338	1.224	0.024
		RF	**0.300**	**0.419**	0.500	0.595	1.813	2.617	0.022
	Chat-Haruhi	EF	**0.303**	**0.433**	**0.505**	0.644	**1.870**	**0.503**	**0.001**
		ER	0.280	0.405	0.500	0.568	2.379	2.400	0.043
		RF	0.267	0.403	0.472	**0.663**	2.109	1.317	0.013
GPT3.5	None	EF	0.317	0.448	0.514	0.665	1.823	**-0.459**	0.020
		ER	0.324	0.454	**0.552**	0.570	1.816	-0.611	0.004
		RF	**0.328**	**0.466**	0.526	**0.697**	**1.550**	-0.778	**0.006**
	RoleLLM	EF	0.290	**0.425**	0.483	**0.699**	1.983	**1.117**	0.003
		ER	**0.301**	0.424	**0.505**	0.603	1.947	2.926	0.009
		RF	0.290	0.414	0.484	0.639	**1.575**	2.306	**0.001**
	Chat-Haruhi	EF	**0.310**	**0.426**	0.486	**0.659**	1.710	**1.352**	0.004
		ER	0.225	0.383	0.469	0.648	2.067	3.807	**0.001**
		RF	0.298	0.422	**0.492**	0.639	**1.659**	1.900	0.003
GPT4o-mini	None	EF	**0.321**	**0.436**	0.502	**0.635**	**1.364**	-0.566	0.006
		ER	0.260	0.397	**0.535**	0.498	2.292	**0.067**	0.024
		RF	0.313	0.435	0.517	0.598	1.658	-0.496	**0.004**
	RoleLLM	EF	**0.285**	**0.414**	0.475	**0.679**	**1.838**	**1.790**	**0.006**
		ER	0.224	0.362	**0.478**	0.543	3.179	4.410	0.009
		RF	0.262	0.387	0.467	0.610	2.159	4.216	0.011
	Chat-Haruhi	EF	**0.266**	**0.394**	**0.476**	**0.610**	**1.867**	**2.032**	**0.000**
		ER	0.186	0.330	0.440	0.567	3.629	4.503	0.030
		RF	0.238	0.367	0.457	0.589	2.247	2.314	0.010

Table 12: Evaluation results of GPT series models on PELD-multi. **EF** stands for "Emotion First", **ER** for "Emotion and Respond", and **RF** for "Respond First". **Error rate** indicates the proportion of instances where model's response contains emotions outside predefined range or irrelevant content.

Model	Type		Eomtion Consistency				EDD	RCD	Error rate
			lower	EC	upper	REC			
LLama3-70b	None	EF	**0.309**	**0.437**	**0.500**	**0.670**	1.822	-1.265	0.060
		ER	0.234	0.375	0.476	0.583	2.936	**0.002**	0.130
		RF	0.293	0.415	0.479	0.656	**1.681**	-0.413	**0.012**
	RoleLLM	EF	**0.243**	**0.402**	**0.472**	**0.694**	**2.399**	**1.360**	0.066
		ER	0.239	0.373	0.471	0.578	3.014	3.457	0.229
		RF	0.203	0.343	0.428	0.622	2.869	4.948	**0.027**
	Chat-Haruhi	EF	**0.269**	**0.410**	**0.485**	0.653	**2.010**	**0.507**	**0.029**
		ER	0.240	0.374	0.466	0.593	2.969	3.124	0.258
		RF	0.210	0.360	0.437	**0.661**	2.653	4.974	0.081
LLama3-8b	None	EF	0.235	0.416	0.507	**0.665**	2.637	-1.777	**0.014**
		ER	0.274	**0.438**	**0.554**	0.586	2.909	-0.440	0.283
		RF	**0.303**	0.436	0.509	0.646	**1.845**	**-0.288**	**0.014**
	RoleLLM	EF	0.185	0.382	0.461	**0.714**	3.141	**1.615**	**0.010**
		ER	**0.281**	**0.440**	**0.537**	0.621	3.153	5.667	0.281
		RF	0.224	0.382	0.462	0.664	**2.511**	4.037	0.017
	Chat-Haruhi	EF	**0.284**	**0.430**	**0.523**	0.611	**1.942**	**3.322**	**0.013**
		ER	0.256	0.418	0.520	**0.614**	3.309	7.953	0.305
		RF	0.248	0.389	0.479	0.610	2.479	7.826	0.014

Table 13: Evaluation results of Llama series models on PELD-multi.

Model	Type		Eomtion Consistency				EDD	RCD	Error rate
			lower	EC	upper	REC			
Doubao-Character	None	EF	0.227	0.374	0.480	**0.581**	2.403	**-0.009**	**0.004**
		ER	**0.288**	**0.423**	**0.535**	0.547	2.420	0.206	0.018
		RF	0.275	0.408	0.531	0.520	**2.178**	-0.772	0.012
	RoleLLM	EF	0.226	0.381	0.490	**0.587**	2.422	**-0.421**	**0.001**
		ER	**0.254**	**0.386**	**0.503**	0.530	2.744	2.488	0.044
		RF	0.246	0.384	0.490	0.566	**2.187**	1.094	0.009
	Chat-Haruhi	EF	**0.257**	**0.391**	0.497	**0.558**	**2.009**	**0.524**	**0.000**
		ER	0.246	0.372	**0.501**	0.494	2.794	1.469	0.009
		RF	0.244	0.381	0.497	0.542	2.090	1.043	0.022
Westlake-7b	None	EF	**0.330**	0.457	0.517	0.679	**1.314**	-0.812	0.148
		ER	0.329	**0.471**	**0.545**	0.657	2.099	**-0.234**	**0.145**
		RF	0.312	0.440	0.494	**0.703**	1.729	-1.507	0.150
	RoleLLM	EF	**0.314**	0.434	0.483	**0.710**	**1.357**	0.508	**0.109**
		ER	0.313	**0.462**	**0.549**	0.631	2.644	**0.471**	0.139
		RF	0.260	0.384	0.452	0.646	2.039	2.102	0.150
	Chat-Haruhi	EF	0.302	0.419	0.480	**0.657**	**1.447**	1.677	**0.060**
		ER	**0.320**	**0.460**	**0.536**	0.648	2.433	1.225	0.118
		RF	0.252	0.377	0.459	0.604	2.211	**1.099**	0.098

Table 14: Evaluation results of role-playing LLMs on PELD-multi. **EF** stands for "Emotion First", **ER** for "Emotion and Respond", and **RF** for "Respond First". **Error rate** indicates the proportion of instances where model's response contains emotions outside predefined range or irrelevant content.

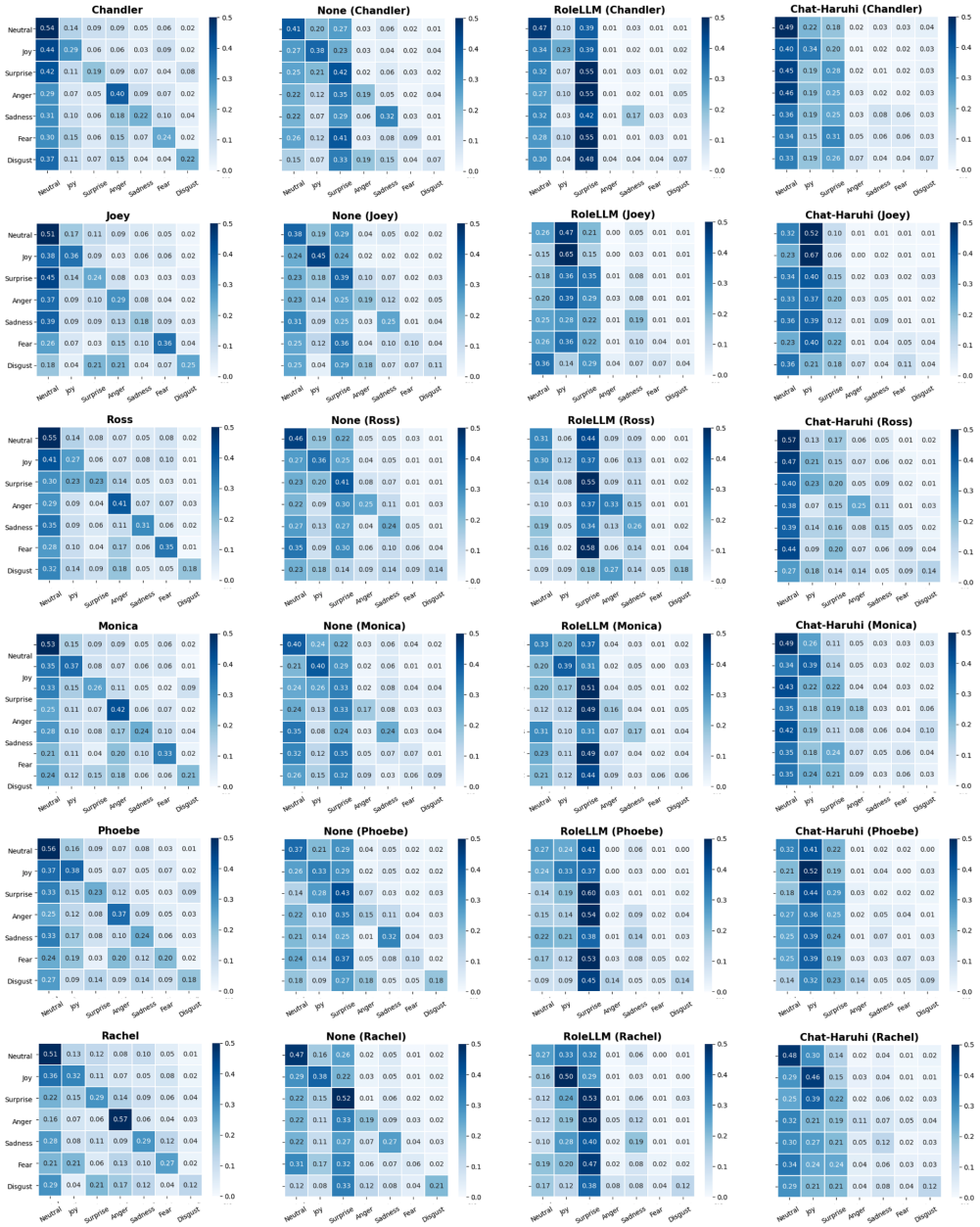

Figure 15: Emotional transition matrices of six characters for different role-playing methods on GPT-3.5.

Joey:	You liked it? You really liked it? **(surprise 😮)**	U_{gt}^1
Chandler:	Oh-ho-ho, yeah! **(joy 😄)**	U_{gt}^2
Joey:	Which part exactly? **(neutral 😐)**	U_{gt}^3
Chandler:	The whole thing! Can we go? **(neutral 😐)**	U_{gt}^4
Joey:	Oh no-no-no, give me some specifics. **(anger 😡)**	U_{gt}^5
Chandler:	I love the specifics, the specifics were the best part! **(joy 😄)**	U_{gt}^6
Joey:	Hey, what about the scene with the kangaroo? Did-did you like that part? **(neutral 😐)**	U_{gt}^7
Chandler:	I was surprised to see a kangaroo in a World War I epic. **(surprise 😮)**	U_{gt}^8
Joey:	You fell asleep! **(anger 😡)**	U_{gt}^9

Figure 16: A multi-turn dialogue case in PELD-multi.