# SELF-EMO: SELF-EVOLVING MULTI-TASK LEARNING FOR EMOTION RECOGNITION IN CONVERSATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Emotion recognition in conversation (ERC) is central to emotionally coherent interaction, yet most systems learn from fixed annotated dialogues. We propose **SELF-EMO**, based on the following hypothesis: better recognition of another speaker's emotion can support a more certain response emotion and a more coherent reply. The model generates three outputs in sequence: an interlocutor emotion, a response emotion, and a reply. SELF-EMO uses these outputs in a training loop that generates, selects, and reuses new dialogue turns. A weighted IoU reward scores ERC predictions, while **SELF-GRPO** encourages agreement within each rollout group. During training, policy and auxiliary-output entropy decline overall. Later analyses and replies receive higher preference from two LLM judges, and a small human study finds more plausible analyses after training. We train one model on pooled IEMOCAP, MELD, and EmoryNLP data. Under this protocol, SELF-EMO obtains the highest dataset-average results among the compared systems. It improves accuracy by 6.32 points for Qwen3-4B and 8.54 points for Qwen3-8B. Weighted F1 improves by 6.33 and 7.83 points. External models construct the initial training data, but final ERC inference requires only the dialogue and candidate emotion labels.

## 1 INTRODUCTION

Emotion recognition in conversation (ERC) predicts a speaker's emotion from the current utterance and its dialogue history. The same utterance can convey different emotions in different contexts, and a speaker's state can change as the conversation unfolds. Recurrent networks, graph models, and pretrained language models have all been used to capture these dependencies (Zhang et al., 2019; Majumder et al., 2019; Ghosal et al., 2019; Shen et al., 2021; Qin et al., 2023). LLMs offer a flexible interface for contextual reasoning, but ERC training still depends on a limited set of labeled dialogues.

Recent LLM-based systems enrich ERC inputs with demonstrations, speaker biographies, or generated descriptions (Lei et al., 2024; Xue et al., 2024; Li et al., 2025a). Such auxiliary information is commonly constructed for a fixed corpus and consumed as static context. We take a different approach. Starting from a labeled dialogue, we generate new role-conditioned continuations, score them with the ERC label, and reuse the best candidates for training. The method still relies on human ERC labels, but the generated context becomes training data instead of an additional inference-time input. Figure 1 illustrates this distinction.

The design follows an observation from psychology: how a person interprets another speaker's emotion affects the person's own reaction and reply (Lane & Schwartz, 1987; Fonagy et al., 2018; van der Schalk et al., 2015). SELF-EMO represents this process with three outputs: the interlocutor emotion, a generated affect label for the next reply, and the reply. We call the second output the *response emotion*. These outputs share one policy and are generated in sequence. We therefore hypothesize that successful ERC optimization can reduce uncertainty across all three outputs.

SELF-EMO uses the official training splits of three ERC datasets. Seed-1.8 and DeepSeek-V3.2 extract personas and scenarios and generate the initial structured examples. The training mixture also contains 15,000 randomly sampled original ERC examples in their standard single-label format. During reinforcement learning, the policy samples several outputs for each prompt. A weighted intersection-over-union (IoU) reward scores the ERC prediction, and SELF-GRPO adds a reward for



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

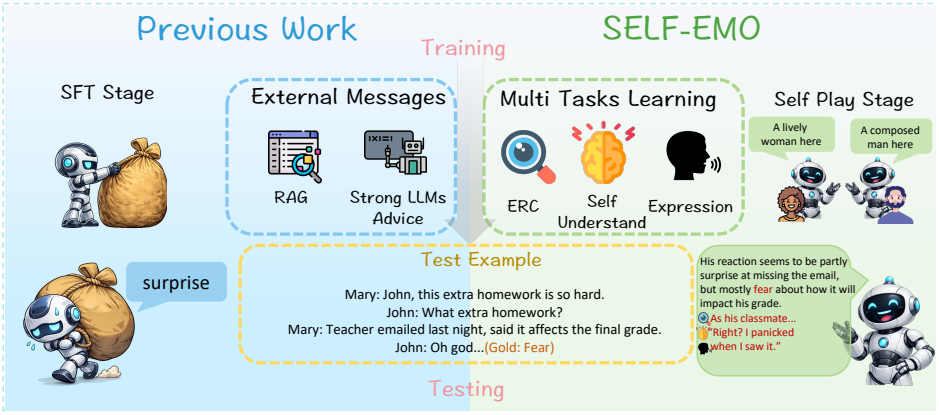



Figure 1: Conceptual comparison between static external-context ERC pipelines and the generate, filter, and reuse training paradigm of SELF-EMO. The structured outputs shown on the right are used during training; final benchmark inference returns only the ERC label.

agreement within the rollout group. The reply from the highest-scoring rollout extends the dialogue and enters the replay buffer. This process alternates reinforcement learning with supervised training. External models are needed to construct the training data, but not at inference time.

We train one model on the pooled training splits of IEMOCAP, MELD, and EmoryNLP and evaluate it separately on each official test split. The test prompt contains the dialogue and the candidate labels for that dataset. It contains no persona, and the model returns one label. SELF-EMO achieves the strongest dataset-average results among the systems evaluated under this pooled protocol. It raises accuracy from 52.97 to 59.29 for Qwen3-4B and from 53.43 to 61.97 for Qwen3-8B. The entropy curves show an overall decline across the shared policy and auxiliary outputs. Two external judges also prefer the analyses and replies produced by later checkpoints.

Our contributions are threefold:

- We develop an ERC training pipeline that expands labeled dialogues with selected role-conditioned continuations and reuses them for training.
- We introduce SELF-GRPO, which combines weighted multi-label feedback with a group agreement reward.
- We show that rising ERC reward coincides with lower entropy in the shared policy and in both auxiliary output fields. Two LLM judges and human raters also prefer the later outputs. SELF-EMO obtains the highest average score under our pooled evaluation protocol.

## 2 Related Work

### 2.1 Emotion Recognition in Conversation

Early ERC systems track contextual and speaker-specific states with recurrent architectures. DialogueRNN models speaker states over a dialogue, while DialogueCRN introduces contextual reasoning modules (Majumder et al., 2019; Hu et al., 2021a). Graph-based methods represent utterances and their relations explicitly. DialogueGCN connects contextual utterances, and DAG-ERC encodes dialogue structure with a directed acyclic graph (Ghosal et al., 2019; Shen et al., 2021). Other systems incorporate commonsense information, multimodal relations, or curriculum learning (Ghosal et al., 2020; Hu et al., 2021b; Nguyen et al., 2024). Pretrained-language-model methods such as BERT-ERC instead adapt contextual encoders directly to emotion classification (Qin et al., 2023).

### 2.2 LLM-BASED AND PERSONA-AUGMENTED ERC

LLM-based ERC methods often cast classification as instruction following and add retrieved or generated context. InstructERC combines instructions, retrieval, and auxiliary objectives (Lei et al., 2024). BiosERC adds speaker biographies (Xue et al., 2024), and PRC-Emo uses internal and external emotion descriptions with curriculum learning (Li et al., 2025a). SELF-EMO also uses persona and scenario descriptions, but for a different purpose: they condition new dialogue continuations that are scored and returned to training. The external LLMs are therefore part of training-data construction.

### 2.3 PSYCHOLOGICAL MOTIVATION FOR COUPLED EMOTION OUTPUTS

Psychological studies connect emotion recognition, appraisal, and response selection. Emotional awareness begins with identifying and differentiating affective states (Lane & Schwartz, 1987). Mentalization then uses such interpretations to guide social reasoning (Fonagy et al., 2018). Social appraisal studies also show that another person's emotion can shape an observer's reaction and decision (van der Schalk et al., 2015). These findings motivate our three-output design. We use them as design motivation rather than as a claim about how an LLM processes emotion.

### 2.4 SELF-TRAINING AND SELF-PLAY

Self-play and synthetic-data methods train models on tasks or trajectories generated during optimization (Fang et al., 2026; Zhang et al., 2025; Yang et al., 2025b). Absolute Zero generates reasoning problems and verifies their solutions (Zhao et al., 2025); related approaches synthesize problems for mathematical and code reasoning (Liang et al., 2025; Yao et al., 2024). Emotional replies lack an equivalent executable verifier. SELF-EMO instead uses ERC labels to score each rollout and treats the selected response emotion as the label of the next turn. The method is thus a reward-filtered form of iterative self-training.

## 3 METHOD

Figure 2 gives an overview of the complete training loop. External models first construct persona-conditioned cold-start examples. The policy then produces three structured outputs, receives task and group-consistency feedback, and returns the selected continuation to the replay buffer for subsequent supervised training.

### 3.1 TASK FORMULATION

SELF-EMO is trained with two prompt formats. The standard ERC format contains a dialogue context $C$ and the candidate labels $\mathcal{E}_d$ of the source dataset $d$. It asks the model to return one label:

$$\widehat{y} \sim \pi_\theta(\cdot \mid C, \mathcal{E}_d), \qquad \widehat{y} \in \mathcal{E}_d. \tag{1}$$

We randomly sample 15,000 original training examples in this format. The same format is used at test time, without persona information. Each benchmark keeps its own label inventory, so no cross-dataset mapping is needed.

The multi-output format is used to generate new dialogue turns. It adds a persona description $P$ and asks the policy to sample a sequence $z \sim \pi_\theta(\cdot \mid C, P, \mathcal{E}_d)$. The sequence is parsed as

$$(\mathbf{p}_o, \mathbf{p}_s, u) = \mathrm{parse}(z), \tag{2}$$

where $\mathbf{p}_o$ predicts the interlocutor emotion, $\mathbf{p}_s$ is the response-emotion distribution, and $u$ is the reply. In the generation template of Table 4 these fields are named `last_emotions`, `my_emotions`, and `my_output`. We use the terms interlocutor emotion, response emotion, and reply throughout the paper. The sequence may also contain a short, observable ERC analysis before these fields. We use this analysis only for auxiliary evaluation; it is not a separate task or reward. The three parsed outputs define tasks $T_o$ (interlocutor emotion), $T_s$ (response emotion), and $T_r$ (reply).

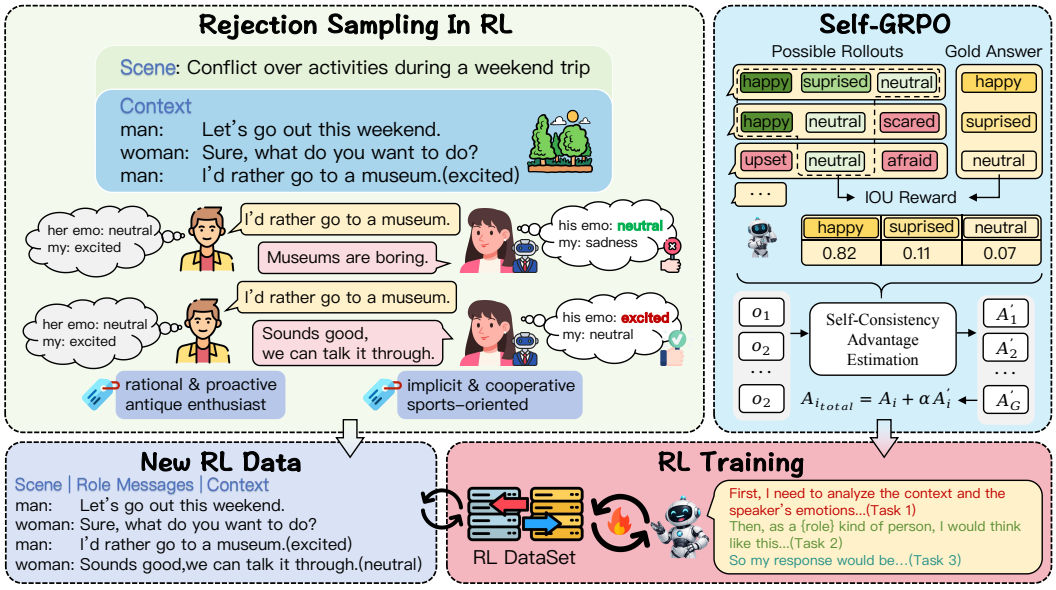



Figure 2: Overview of SELF-EMO. Cold-start construction initializes the structured three-output task. During SELF-GRPO, weighted ERC feedback and within-group agreement optimize the policy; the selected response and response-emotion label extend the replay buffer for subsequent training.

Their autoregressive factorization is

$$\begin{aligned} \mathbf{p}_o &\sim \pi_\theta(\cdot \mid C, P, \mathcal{E}_d), \\ \mathbf{p}_s &\sim \pi_\theta(\cdot \mid C, P, \mathcal{E}_d, \mathbf{p}_o), \\ u &\sim \pi_\theta(\cdot \mid C, P, \mathcal{E}_d, \mathbf{p}_o, \mathbf{p}_s). \end{aligned} \tag{3}$$

The three fields share one policy and appear in the same sequence, so an ERC update can also change the generated analysis and reply.

### 3.2 Empirical Hypothesis of Joint Entropy Reduction

The sequential design couples the three outputs, but it does not impose a deterministic relation among their entropies. We instead formulate a testable hypothesis. When ERC optimization succeeds, uncertainty should decline overall in the shared policy, the response-emotion field, and the reply field.

For a generated position $j$, token entropy is

$$h_j = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid x, y_{<j}) \log \pi_\theta(v \mid x, y_{<j}), \tag{4}$$

where $\mathcal{V}$ is the vocabulary. Lowercase $h$ with a position index denotes this per-token quantity. Uppercase $H$ with a checkpoint index denotes its average at checkpoint $t$. Accordingly, $H_t^\pi$ averages $h_j$ over all generated positions, and $H_t^s$ and $H_t^r$ average $h_j$ over the positions of the $T_s$ and $T_r$ fields. The hypothesis predicts the following overall trend:

$$H_T^\pi < H_0^\pi, \qquad H_T^s < H_0^s, \qquad H_T^r < H_0^r. \tag{5}$$

We evaluate this hypothesis with reward curves, field-level entropy, and pairwise judgments of the generated analyses and replies. Section 7 states what these measurements can and cannot settle.

### 3.3 Persona and Cold-Start Construction

Let $\mathcal{D}_{\text{pool}}$ be the union of the three official training splits. We randomly sample 15,000 standard-format examples from $\mathcal{D}_{\text{pool}}$ and denote them by $\mathcal{D}_{\text{ERC}}$. All data-construction and optimization

steps use training data only. For a training dialogue $c \in \mathcal{D}_{\text{pool}}$, an external model $M_e$ receives prompt $P_{\text{ext}}$ and returns a role profile $P$ and scenario $S$:

$$(P, S) = M_e(c, P_{\text{ext}}). \tag{6}$$

Seed-1.8 and DeepSeek-V3.2 provide these descriptions and generate the initial multi-output examples $\mathcal{D}_{\text{cold}}$, of which we keep 8,000. The two prompt sets therefore hold $|\mathcal{D}_{\text{ERC}}| = 15{,}000$ and $|\mathcal{D}_{\text{cold}}| = 8{,}000$ examples. The scenario summary $S$ is produced only as part of the profile-and-theme extraction step; it is not included in the SFT or reinforcement-learning prompts. Subsequent multi-output generation is conditioned on the role profile $P$, as specified in Eq. 3. Supervised fine-tuning uses both $\mathcal{D}_{\text{ERC}}$ and $\mathcal{D}_{\text{cold}}$. The first set teaches the persona-free single-label task used at test time; the second initializes the multi-output generation used for dialogue expansion.

The role descriptions are generated conditioning text rather than validated personality measurements. Appendix A summarizes the extraction prompt in Table 3, and Section 7 discusses its demographic fields.

### 3.4 Weighted Multi-Label ERC Reward

Each training prompt supplies the candidate labels $\mathcal{E}_d$ of its source dataset. A multi-output rollout may predict up to three labels from this set. Let $Q_i = \{(e, w^Q_{i,e})\}$ denote rollout $i$'s prediction and $Y = \{(e, w^Y_e)\}$ its target. For a standard original example, $Y$ is the human ERC label. External models provide soft multi-label targets for the cold-start structured examples. For a generated continuation, $Y$ is the response-emotion pseudo-label from the preceding selected rollout. We normalize the weights within each set:

$$\widehat{w}^Q_{i,e} = \frac{w^Q_{i,e}}{\sum_{e'} w^Q_{i,e'}}, \qquad \widehat{w}^Y_e = \frac{w^Y_e}{\sum_{e'} w^Y_{e'}}. \tag{7}$$

Unlisted labels receive zero weight. The weighted overlap is

$$R^{\text{IoU}}_i = \frac{\sum_{e \in Q_i \cup Y} \min(\widehat{w}^Q_{i,e}, \widehat{w}^Y_e)}{\sum_{e \in Q_i \cup Y} \max(\widehat{w}^Q_{i,e}, \widehat{w}^Y_e)}. \tag{8}$$

This reward gives partial credit whenever the predicted set overlaps the target set; it does not encode semantic distance between distinct labels. The denominator grows with every extra predicted label. The reward is therefore higher when the predicted mass is concentrated on the target than when several candidates are listed. A format checker $g(y_i)$ validates the structured fields. The primary reward is

$$R^{\text{task}}_i = \begin{cases} 0, & g(y_i) = 0, \\ R^{\text{IoU}}_i + 0.1, & g(y_i) = 1. \end{cases} \tag{9}$$

### 3.5 SELF-GRPO

We extend GRPO (Shao et al., 2024) with an additional within-group label-overlap reward. For a prompt $x$, the old policy samples a group $\{y_i\}_{i=1}^n$. Let $Q_i$ be the emotion-label support in rollout $i$. We aggregate the normalized predicted masses across the group and select the $m = \min(3, |\mathcal{E}_d|)$ labels with largest aggregate mass, denoted $\mathcal{E}_{\text{core}}$. The secondary reward measures overlap with this support:

$$R^{\text{grp}}_i = \frac{1}{m} \sum_{e \in \mathcal{E}_{\text{core}}} \mathbb{I}[e \in Q_i]. \tag{10}$$

This term rewards agreement rather than correctness, so its preference differs from that of Eq. 8. Original examples use human ERC labels, whereas generated continuations use response-emotion pseudo-labels. The latter can reinforce an earlier error. We therefore raise $\lambda_\tau$ gradually instead of applying its full weight from the start.

For a reward $R$ within a rollout group, define

$$\text{norm}(R_i) = \frac{R_i - \mu_R}{\sigma_R + \epsilon_{\text{norm}}}, \tag{11}$$

**Algorithm 1** SELF-EMO Closed-Loop Optimization

**Require:** Standard ERC data $\mathcal{D}_{\text{ERC}}$; cold-start data $\mathcal{D}_{\text{cold}}$; outer iterations $K$; rollouts per prompt $n$; rollout batch size $B$; mini-batch size $b$

Initialize $\pi_\theta$ by SFT on $\mathcal{D}_{\text{ERC}} \cup \mathcal{D}_{\text{cold}}$
Initialize replay buffer $\mathcal{B}$ with the training trajectories
**for** $k = 1, \ldots, K$ **do**
  **for** each batch $\mathcal{X}$ of $B$ prompts drawn from $\mathcal{B}$ **do**
    $\theta_{\text{old}} \leftarrow \theta$
    **for** each $x \in \mathcal{X}$ **do**
      Sample $\{y_i\}_{i=1}^{n} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$
      Compute $R_i^{\text{task}}$, $R_i^{\text{grp}}$, and $A_i$
    **end for**
    **for** each mini-batch of $b$ prompts in $\mathcal{X}$ **do**
      $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{SELF-GRPO}}$
    **end for**
    **for** each multi-output prompt $x \in \mathcal{X}$ **do**
      Select $y^* = \arg\max_i R_i^{\text{task}}$
      Construct a continuation with Eq. 15 and add it to $\mathcal{B}$
    **end for**
  **end for**
  Continue supervised training of $\pi_\theta$ on the updated $\mathcal{B}$
**end for**
**return** $\pi_\theta$

where $\epsilon_{\text{norm}} > 0$ prevents division by zero. At optimization step $\tau$ out of $T_{\max}$, the advantage is

$$A_i = \text{norm}(R_i^{\text{task}}) + \lambda_\tau \, \text{norm}(R_i^{\text{grp}}), \qquad \lambda_\tau = \frac{\tau}{T_{\max}}. \tag{12}$$

We use $\tau$ here to distinguish the optimization step from the checkpoint index $t$ in Section 3.2. Let $y_i = (y_{i,1}, \ldots, y_{i,|y_i|})$. At token position $j$, the importance ratio is

$$\rho_{i,j}(\theta) = \frac{\pi_\theta(y_{i,j} \mid x, y_{i,<j})}{\pi_{\theta_{\text{old}}}(y_{i,j} \mid x, y_{i,<j})}. \tag{13}$$

The rollout-level advantage $A_i$ is shared by its generated tokens. Let $d_{i,j}^{\text{KL}}$ be the token-level low-variance KL estimate between the policy and the reference model. We minimize

$$\mathcal{L}_{\text{SELF-GRPO}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|y_i|}\sum_{j=1}^{|y_i|}\Big[\min\left(\rho_{i,j}A_i, \text{clip}(\rho_{i,j}, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})A_i\right) - \beta d_{i,j}^{\text{KL}}\Big]. \tag{14}$$

We use $\beta = 0.001$ and $\epsilon_{\text{clip}} = 0.2$. The KL term is part of the actor loss and is not added to the reward.

### 3.6 Generate, Filter, and Reuse Training

For each prompt, we retain the rollout with the highest task reward. Its reply $u^*$ is appended to the dialogue, and its response emotion becomes the ERC pseudo-label of the new turn:

$$C_{\text{new}} = C \parallel u^*, \qquad Y_{\text{new}} = \mathbf{p}_s^*. \tag{15}$$

The new turn is added to the replay buffer. SELF-GRPO and supervised training then alternate on the updated data, as shown in Algorithm 1. The task reward directly scores $\mathbf{p}_o$, while the replay step reuses $\mathbf{p}_s$ and $u$. We therefore assess auxiliary generation quality separately in Section 5. The LLM judges prefer the analyses and replies from later checkpoints during the same training period in which ERC reward rises.

# 4 EXPERIMENTS

## 4.1 DATA USE AND EVALUATION PROTOCOL

We use IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and EmoryNLP (Zahiri & Choi, 2018). Their train/development/test sizes are 5,163/647/1,623, 9,989/1,109/2,610, and 7,551/954/984, respectively. We pool the three training splits and randomly sample 15,000 original examples for the standard ERC branch. Persona extraction, synthetic-data generation, reinforcement learning, and replay use training data only. Periodic evaluation uses a held-out diagnostic subset derived from the training data. The official test sets are used only for final evaluation.

The final prompt contains the dialogue and the candidate emotions for the corresponding dataset. It contains no persona or scenario, and the model returns one label. This is the same format used by the 15,000 original ERC examples during training. Following the evaluation practice of prior ERC studies, we report Accuracy and weighted F1 (W-F1), computed directly from the returned labels. Because each prompt provides its own candidate set, the three datasets can keep their original label inventories.

## 4.2 COMPARED SYSTEMS AND IMPLEMENTATION

Table 1 first lists dataset-specific results reported in prior work. They provide context but are not directly comparable because the models, modalities, and training protocols differ. The lower block uses unified testing: every system is evaluated on the same three official test splits with the same metrics and the candidate-label inventory of each dataset. It includes direct evaluations of Seed-1.8 and DeepSeek-V3.2, untuned Qwen3 references, two pooled retraining baselines, and SELF-EMO. Inference-time auxiliary inputs remain method-specific, as indicated by the "Inf. Aux." column. BiosERC and PRC-Emo were originally designed for dataset-specific training and rely on static, dataset-specific auxiliary knowledge or profiles. One model is trained jointly here across three dialogue domains and label inventories. We therefore attribute part of their lower pooled results to the limited transfer of this fixed information.

SELF-EMO uses Qwen3-4B and Qwen3-8B (Yang et al., 2025a). We train on one node with eight NVIDIA H100 GPUs using VeRL. Each prompt has eight rollouts, the rollout batch size is $B = 64$ prompts, the mini-batch size is $b = 32$, and the actor learning rate is $10^{-6}$. We run $K = 3$ outer iterations, which is the `trainer.total_epochs` setting in Appendix F. Each iteration appends one new turn per multi-output prompt. The replay buffer therefore grows by about 8k examples per iteration and holds roughly 32k multi-output examples at the end of training. The full Qwen3-8B run takes about 310 H100 GPU-hours. Appendix F gives the remaining settings. Seed-1.8 and DeepSeek-V3.2 construct training data and serve as auxiliary judges. They are not used during final ERC inference. Results produced by our pooled-training runs are means over four random seeds. We do not report per-seed dispersion or significance tests.

## 4.3 ERC RESULTS UNDER UNIFIED TESTING

SELF-EMO-4B scores above Qwen3-4B on all three test sets. Dataset-average accuracy rises from 52.97 to 59.29, a gain of 6.32 points. For the 8B model, it rises from 53.43 to 61.97, a gain of 8.54 points. The corresponding dataset-average W-F1 gains are 6.33 and 7.83 points. The largest 8B accuracy gain occurs on MELD (12.49 points), followed by IEMOCAP (8.44) and EmoryNLP (4.68). SELF-EMO-8B has the highest dataset-average Accuracy and W-F1 in the unified-testing block of Table 1. This is the best result under our pooled evaluation protocol.

## 4.4 COMPONENT ABLATIONS

SELF-GRPO is 3.13 points above vanilla GRPO when both use 8k synthetic and 15k original examples. SELF-EMO is also 2.14 points above the expert-SFT baseline, despite using fewer synthetic examples. This comparison shows that adding expert-generated SFT data alone does not explain the gain from selected self-generated data. Training only on the original data gives 54.12 average accuracy, 5.17 points below SELF-EMO.

| Method | Backbone | Inf. Aux. | IEMOCAP Acc. | IEMOCAP W-F1 | MELD Acc. | MELD W-F1 | EmoryNLP Acc. | EmoryNLP W-F1 | Average Acc. | Average W-F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset-specific | | | | | | | | | | |
| DialogueRNN (Majumder et al., 2019) | GRU | – | 63.40 | 62.75 | 60.27 | 57.03 | – | 31.70 | – | 50.49 |
| ICON (Hazarika et al., 2018) | – | – | 64.00 | 63.50 | – | 56.30 | – | – | – | – |
| DialogueGCN (Ghosal et al., 2019) | – | – | 65.25 | 64.18 | – | 58.10 | – | – | – | – |
| COSMIC (Ghosal et al., 2020) | – | – | – | 65.28 | – | 65.21 | – | 38.11 | – | 56.20 |
| MMGCN (Hu et al., 2021b) | – | – | – | 66.22 | – | 58.65 | – | – | – | – |
| DAG-ERC (Shen et al., 2021) | – | – | 67.53 | 68.03 | 63.98 | 63.63 | – | 39.02 | – | 56.89 |
| LR-GCN (Ren et al., 2022) | – | – | 68.50 | 68.30 | – | 65.60 | – | – | – | – |
| MultiDAG+CL (Nguyen et al., 2024) | – | – | 69.11 | 69.08 | – | 64.00 | – | – | – | – |
| CBERL (Meng et al., 2023) | – | – | 69.36 | 69.27 | 67.78 | 66.89 | – | – | – | – |
| DER-GCN (Ai et al., 2024) | – | – | 69.70 | 69.40 | 66.80 | 66.10 | – | – | – | – |
| LSDGNN+ICL (Li et al., 2025b) | – | – | 70.35 | 70.24 | 64.67 | 64.07 | – | – | – | – |
| InstructERC (Lei et al., 2024) | LLaMA2-7B | Yes | – | 71.39 | – | 69.15 | – | 41.39 | – | – |
| BiosERC (Xue et al., 2024) | LLaMA2-13B | Yes | – | 71.19 | – | 69.83 | – | 41.68 | – | – |
| PRC-Emo (Li et al., 2025a) | Qwen2.5/3 | Yes | 71.03 | 71.95 | 71.50 | 70.44 | – | – | – | – |
| Unified testing | | | | | | | | | | |
| Seed-1.8 | Seed-1.8 | No | 18.30 | 18.54 | 26.82 | 25.31 | 21.85 | 20.99 | 22.32 | 21.61 |
| DeepSeek-V3.2 | DeepSeek-V3.2 | No | 19.90 | 19.84 | 20.69 | 20.57 | 19.82 | 18.55 | 20.14 | 19.65 |
| Qwen3 | 4B | No | 56.07 | 55.83 | 60.96 | 59.55 | 41.87 | 41.50 | 52.97 | 52.29 |
| Qwen3 | 8B | No | 57.67 | 57.33 | 59.43 | 59.21 | 43.19 | 42.71 | 53.43 | 53.08 |
| BiosERC | LLaMA2-13B | Yes | 28.53 | 28.84 | 62.13 | 61.59 | 19.83 | 20.55 | 36.83 | 36.99 |
| PRC-Emo | Qwen3-8B | Yes | 27.39 | 27.53 | 64.07 | 63.88 | 22.53 | 21.94 | 38.00 | 37.78 |
| SELF-EMO | Qwen3-4B | No | **64.60** | **63.57** | **67.74** | **67.32** | **45.52** | **44.98** | **59.29** | **58.62** |
| SELF-EMO | Qwen3-8B | No | **66.11** | **64.81** | **71.92** | **70.30** | **47.87** | **47.61** | **61.97** | **60.91** |

Table 1: ERC results on IEMOCAP, MELD, and EmoryNLP. "Dataset-specific" reports values from the cited papers, where each dataset is trained and evaluated separately under heterogeneous protocols. "Unified testing" evaluates all listed systems on the same three official test splits with the same metrics and each dataset's candidate-label inventory. In this block, Seed-1.8, DeepSeek-V3.2, and untuned Qwen3 are single evaluations, whereas the pooled-trained BiosERC, PRC-Emo, and SELF-EMO rows are four-seed means. "Inf. Aux." indicates whether retrieved or generated auxiliary context is required at final inference; SELF-EMO uses external models only for training-data construction.

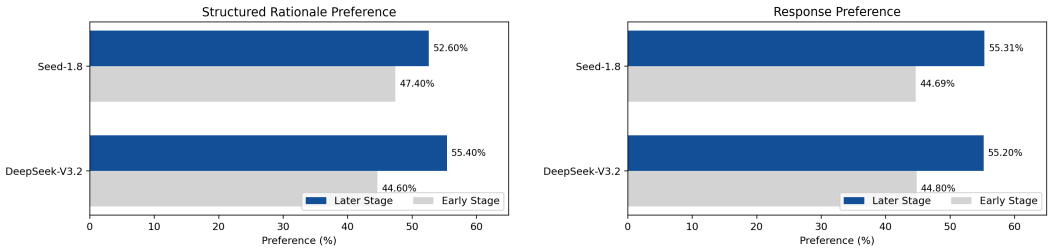
| Variant | Synth. | Orig. | Emory | IEMOCAP | MELD | Avg. |
|---|---|---|---|---|---|---|
| SELF-EMO | 8k | 15k | 45.52 | 64.60 | 67.74 | 59.29 |
| Expert-SFT + GRPO | 13k | 15k | 42.46 | 61.95 | 67.03 | 57.15 |
| Vanilla GRPO | 8k | 15k | 42.23 | 61.10 | 65.14 | 56.16 |
| Original-data GRPO | 0k | 22.5k | 42.07 | 58.15 | 62.13 | 54.12 |

Table 2: Qwen3-4B ablations. All configurations use the same output-format initialization. The expert-SFT baseline then uses 13k expert-generated examples before RL, whereas SELF-EMO trains with 8k selected self-generated examples and 15k randomly sampled original ERC examples.

# 5 Analysis

## 5.1 Auxiliary-Output Preference Analysis

The ERC reward directly scores the interlocutor-emotion field, not the generated analysis or reply. We therefore compare these outputs at an early and a later RL checkpoint on a random sample of 500 dialogue examples. Seed-1.8 and DeepSeek-V3.2 receive anonymized pairs and choose the output that better fits the dialogue and role description. In Figure 3, blue denotes the later checkpoint and gray the early checkpoint. Both judges favor the later analyses and replies. Together with the rising ERC reward in Figure 5, this result supports joint improvement of the optimized ERC task and the two generated outputs during training. Appendix D gives the judging protocol and paired examples.




(a) Generated-analysis preference.

(b) Response preference.

Figure 3: External-model A/B preferences between later- and early-stage outputs on 500 randomly sampled dialogue examples. Blue denotes Later Stage and gray denotes Early Stage.

### 5.2 Joint Entropy Reduction

This analysis tests the hypothesis in Eq. 5. Both the benchmark reward and the weighted-IoU reward rise during training (Appendix C). Policy entropy falls quickly at first and then declines more slowly. Response-emotion entropy follows the same overall direction. Reply-token entropy drops early, with a small rebound near the end. All three final entropy values remain below their initial values, which matches Eq. 5. The pairwise judge study provides a separate quality check: both judges prefer the analyses and replies from later checkpoints.

### 5.3 Results by Persona Cluster

HDBSCAN groups the generated persona descriptions into 74 clusters. For the five most frequent clusters, SELF-EMO obtains mean rewards from 56.32% to 61.98%, with a five-cluster average of 58.48%. Seed-1.8 (13.77%) and DeepSeek-V3.2 (10.04%) obtain lower rewards under the same evaluation, although both are scored on the objective that SELF-EMO was trained on. Appendix B gives the cluster definitions and the full distribution. These clusters summarize generated descriptions and are not psychological measurements.

A small human study and an eight-rollout stability analysis show the same direction. Human raters find SELF-EMO's analyses more plausible than Qwen3's by 10 points overall. SELF-EMO also improves single-rollout accuracy and prediction consistency, while its Oracle@8 decreases. Appendix E reports both studies.

## 6 Discussion

SELF-EMO uses ERC supervision twice. It trains on 15,000 original single-label examples, which matches the final inference format, and it uses the same labels to select multi-output rollouts that become new dialogue turns. The ablations separate the two effects. Under the same 8k synthetic and 15k original examples, SELF-GRPO exceeds vanilla GRPO by 3.13 points, and SELF-EMO exceeds the expert-SFT baseline by 2.14 points with fewer synthetic examples. Expert-generated SFT data alone therefore does not explain the gain.

The remaining analyses agree with this result. ERC reward rises while policy, response-emotion, and reply-token entropy decline. Both judges prefer later checkpoints, human raters find the analyses more plausible, and single-rollout accuracy and consistency improve.

The group reward remains the main open risk. Agreement within a group makes the structured output easier to learn, but it can also reinforce a shared mistake. Replayed turns also use model-generated response emotions instead of human labels. Future work should measure pseudo-label accuracy and class-wise diversity across replay depth.

## 7 LIMITATIONS

We train on three known ERC corpora and do not test transfer to an unseen dataset, another language, or real user conversations. MELD and EmoryNLP come from television dialogue and IEMOCAP from acted interactions.

The entropy result is an observed pattern rather than a guarantee. It need not hold at every checkpoint, lower entropy alone does not mean better output, and we report no sample-level correlation between ERC reward and auxiliary-output quality. In Table 2, the original-data row uses 22.5k rather than 15k original examples, so it is not a clean one-factor comparison. Reported numbers are four-seed means, but the tables omit dispersion. The checkpoint comparison relies on LLM judges. The human study is small and rates analysis plausibility rather than replay-label accuracy or reply quality.

Seed-1.8 and DeepSeek-V3.2 both build the training data and judge the outputs, and their corpora and service versions are not observable. Replayed turns use generated response emotions, so an incorrect pseudo-label can propagate.

The persona prompt includes estimated age, gender, and educational background. These inferences can reproduce stereotypes and are not needed for ERC, so future versions should keep only observable speaking style. The generated profiles should not be used in clinical, employment, education, or surveillance settings.

## 8 CONCLUSION

SELF-EMO connects emotion recognition, response-emotion generation, and reply generation in one iterative training loop. It uses a weighted ERC reward to select new dialogue turns and SELF-GRPO to encourage agreement within each rollout group. ERC reward rises, the three entropy measures decline overall, and two LLM judges prefer the outputs of later checkpoints. Human ratings and the rollout study show the same direction. Under the pooled protocol, SELF-EMO obtains the highest average score among the compared systems. Final ERC inference requires only the dialogue and the dataset-specific candidate labels.

### AI USE STATEMENT

Generative AI tools assisted with language editing, structural revision, and consistency checks. The authors verified the equations, claims, references, tables, and figures and take responsibility for the final manuscript.

### ETHICS STATEMENT

SELF-EMO generates role descriptions, affect labels, and dialogue continuations. These outputs may contain stereotypes, factual errors, or inappropriate language inherited from the base and external models. The original persona prompt includes coarse demographic inferences, which can be unreliable and harmful. Future data construction should use observable communication style instead. The term "self-emotion" denotes a generation variable, not subjective experience. The method is intended for conversational modeling research and should not be used for diagnosis, eligibility decisions, surveillance, or manipulation. Any data release should follow the licenses of the source corpora and include safety and bias checks.

### REPRODUCIBILITY STATEMENT

Section 3 defines the outputs, rewards, SELF-GRPO objective, and replay procedure. Section 4 describes the data splits and the persona-free test protocol. Appendix A summarizes the prompts, Appendix C reports training curves, and Appendix F lists the VeRL settings. All data-construction and optimization steps use the official training splits. Periodic evaluation uses a held-out subset of the training data, and official test sets are reserved for final evaluation. The main experiments use four random seeds, and the reported results are their means; the current tables omit per-seed dispersion. Some external API decoding settings are not recorded.

## REFERENCES


Wei Ai, Yuntao Shou, Tao Meng, Nan Yin, and Keqin Li. Der-gcn: Dialogue and event relation-aware graph convolutional neural network for multimodal dialogue emotion recognition, 2024. URL https://arxiv.org/abs/2312.10579.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. Serl: Self-play reinforcement learning for large language models with limited data, 2026. URL https://arxiv.org/abs/2505.20347.

Peter Fonagy, Gyorgy Gergely, and Elliot L Jurist. *Affect regulation, mentalization and the development of the self*. Routledge, 2018.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, 2019. URL https://arxiv.org/abs/1908.11540.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations, 2020. URL https://arxiv.org/abs/2010.02795.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICON: Interactive conversational memory network for multimodal emotion detection. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2594–2604, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1280. URL https://aclanthology.org/D18-1280/.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7042–7052, 2021a.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation, 2021b. URL https://arxiv.org/abs/2107.06779.

Richard D Lane and Gary E Schwartz. Levels of emotional awareness: a cognitive-developmental theory and its application to psychopathology. *The American journal of psychiatry*, 144(2):133–143, 1987.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models, 2024. URL https://arxiv.org/abs/2309.11911.

Xinran Li, Yu Liu, Jiaqi Qiao, and Xiujuan Xu. Do llms feel? teaching emotion recognition with prompts, retrieval, and curriculum learning, 2025a. URL https://arxiv.org/abs/2511.07061.

Xinran Li, Xiujuan Xu, and Jiaqi Qiao. *Long-Short Distance Graph Neural Networks and Improved Curriculum Learning for Emotion Recognition in Conversation*. IOS Press, October 2025b. ISBN 9781643686318. doi: 10.3233/faia251292. URL http://dx.doi.org/10.3233/FAIA251292.

Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu Chen. Beyond pass@ 1: Self-play with variational problem synthesis sustains rlvr. *arXiv preprint arXiv:2508.14029*, 2025.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations, 2019. URL https://arxiv.org/abs/1811.00405.

Tao Meng, Yuntao Shou, Wei Ai, Nan Yin, and Keqin Li. Deep imbalanced learning for multimodal emotion recognition in conversations, 2023. URL https://arxiv.org/abs/2312.06337.

Cam-Van Thi Nguyen, Cao-Bach Nguyen, Quang-Thuy Ha, and Duc-Trong Le. Curriculum learning meets directed acyclic graph for multimodal emotion recognition, 2024. URL https://arxiv.org/abs/2402.17269.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 13492–13500, 2023.

Minjie Ren, Xiangdong Huang, Wenhui Li, Dan Song, and Weizhi Nie. Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition. *IEEE Transactions on Multimedia*, 24:4422–4432, 2022. doi: 10.1109/TMM.2021.3117062.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition, 2021. URL https://arxiv.org/abs/2105.12907.

Job van der Schalk, Toon Kuppens, Martin Bruder, and Antony SR Manstead. The social power of regret: the effect of social appraisal and anticipated emotions on fair and unfair allocations in resource dilemmas. *Journal of Experimental Psychology: General*, 144(1):151, 2015.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. *BiosERC: Integrating Biography Speakers Supported by LLMs for ERC Tasks*, pp. 277–292. Springer Nature Switzerland, 2024. ISBN 9783031723445. doi: 10.1007/978-3-031-72344-5_19. URL http://dx.doi.org/10.1007/978-3-031-72344-5_19.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Ziyi Yang, Weizhou Shen, Chenliang Li, Ruijun Chen, Fanqi Wan, Ming Yan, Xiaojun Quan, and Fei Huang. Spell: Self-play reinforcement learning for evolving long-context language models. *arXiv preprint arXiv:2509.23863*, 2025b.

Junyi Yao, Yijiang Liu, Zhen Dong, Mingfei Guo, Helan Hu, Kurt Keutzer, Li Du, Daquan Zhou, and Shanghang Zhang. Promptcot: Align prompt distribution via adapted chain-of-thought. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7027–7037, 2024.

Sayyed M Zahiri and Jinho D Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*, volume 18, pp. 44–52, 2018.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pp. 5415–5421. Macao, 2019.

Ruize Zhang, Zelai Xu, Chengdong Ma, Chao Yu, Wei-Wei Tu, Wenhao Tang, Shiyu Huang, Deheng Ye, Wenbo Ding, Yaodong Yang, and Yu Wang. A survey on self-play methods in reinforcement learning, 2025. URL https://arxiv.org/abs/2408.01072.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.

# A Data-Construction Prompts

This section reproduces the prompts used during cold-start data construction. The scenario/theme field extracted by the first prompt corresponds to $S$ in Section 3. It is retained as extraction metadata and is not passed to the model during subsequent SFT or reinforcement learning. The demographic fields are documented for reproducibility; as discussed in Section 7, future data construction should omit them.

## A.1 Scenario and Persona Extraction

**Scenario and Personality Information Extraction**

You are a professional emotional analyst and personality profiling expert.
Your task is to analyze a given multi-speaker dialogue and produce a distinct personality profile for each participant, as well as extract the overall theme of the conversation.
Please strictly follow the requirements below.
**1. Analysis Dimensions** (Inference-based, no fabrication)
For each speaker, infer their characteristics ONLY based on the dialogue content. The analysis should consider (but is not limited to) the following aspects:
1. Personality traits (e.g., rational vs. emotional, introverted vs. extroverted, cautious vs. impulsive, gentle vs. assertive)
2. Estimated age range (use vague ranges such as "20–30s", "30+"; do NOT provide a specific age)
3. Possible gender (if unclear, explicitly state "gender not obvious")
4. Likely education or cognitive background (e.g., higher-education background, practical/non-academic thinking style)
5. Speaking and expression style (e.g., formal vs. casual, logic-driven vs. emotion-driven, restrained vs. expressive wording)
6. Emotional state and stance (e.g., anxious, complaining, calmly analyzing, comforting others, confrontational vs. cooperative)
Requirements:
- All conclusions must be grounded in the dialogue itself; do NOT invent information.
- Each character's personality profile must be clearly distinctive.
- Avoid generic, vague, or template-like descriptions (e.g., "quite normal", "relatively rational").
**2. Personality Summary Requirement**
- For each speaker, compress the analysis into ONE concise but information-dense sentence.
- The sentence should naturally integrate multiple dimensions rather than listing attributes mechanically.
Example style (for reference only, do NOT copy):
"The speaker appears to be a cautious and rational individual, likely in their 30s, with a higher-education background, expressing themselves in a structured and emotionally restrained manner."
**3. Theme Extraction**
- Abstract one or more core themes from the entire dialogue.
- Themes should be semantic-level summaries, not surface-level restatements.
- If conflicts or disagreements exist, reflect them in the theme when appropriate.
**4. Output Format (Critical)**
After completing the analysis, output a Python code block containing a dictionary in the following format:

```
{
    "Speaker A Name": "One-sentence personality profile",
    "Speaker B Name": "One-sentence personality profile",
    ...
    "Theme": "Core theme(s) of the conversation"
}
```

Output rules:
You should first analyze each speaker, and then output the Python code block.
Do NOT output anything outside the Python code block.
Speaker names must exactly match those used in the dialogue.

Table 3: Prompt used for scenario and personality information extraction.

### A.2 COLD-START STRUCTURED GENERATION

**Prompt for SFT Data Generation**

**System Prompt**
You are now a person named "{name}" with the following characteristics: {role_message}.
You are currently engaged in a conversation with others.
You should:
(1) Infer the emotion, personality traits, and underlying thoughts of the last speaker in the conversation history, and provide an emotional profile (as a dictionary) for that person.
(2) Reflect based on your own role persona, then formulate your response—along with the emotions embedded in your reply.
{emo_list}
**User Prompt**
{history}
You now need to analyze the conversation, provide an analysis of the last utterance along with your own reasoning, and package them into a Python dictionary.
**Assistant Output Format**

```
<think>
... reasoning process ...
</think>

{
 "last_emotions": ...,
 "my_emotions": ...,
 "my_output": ...
}
```

Table 4: Prompt used for cold-start SFT data generation. The fields `last_emotions`, `my_emotions`, and `my_output` correspond to $\mathbf{p}_o$, $\mathbf{p}_s$, and $u$ in Section 3. No generated chain-of-thought is reproduced in this paper.

## B PERSONA-CLUSTER DIAGNOSTICS

We apply HDBSCAN to the extracted persona descriptions. To reduce lexical bias in clustering, original personality names are masked before clustering. This procedure produces 74 clusters. Figure 4 visualizes the twelve largest clusters, and Tables 5 and 6 give the full distribution. Some clusters remain semantically similar because they are induced automatically. Their names summarize patterns in model-generated text and should not be interpreted as verified human traits or diagnoses.

For the five largest clusters in Table 5, we compare the weighted-IoU rewards of SELF-EMO-8B with Seed-1.8 and DeepSeek-V3.2 under the same evaluation. We abbreviate their labels as follows: EDAE for "emotional directness assertive engagement" (1108); EEIPRC for "emotional expressiveness interpersonal playfulness reactive curiosity" (590); HRS for "humor regulated social" (487); CPDE for "composed practical direct efficient" (329); and ERDCPD for "emotional reactivity defensive confrontation pragmatic dominance" (307). Table 7 reports the results in that order.

These generated cluster summaries are not validated psychological scales.

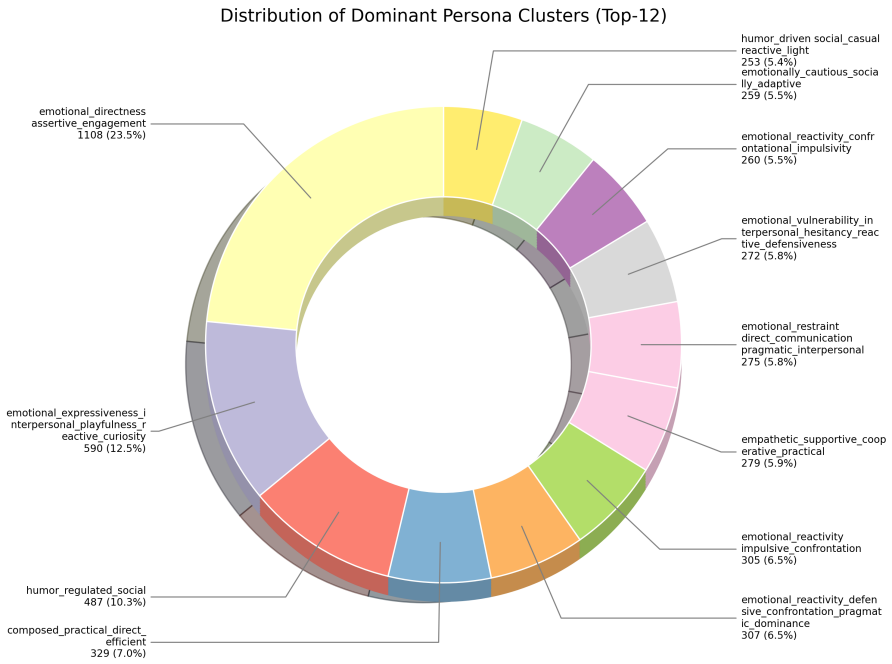
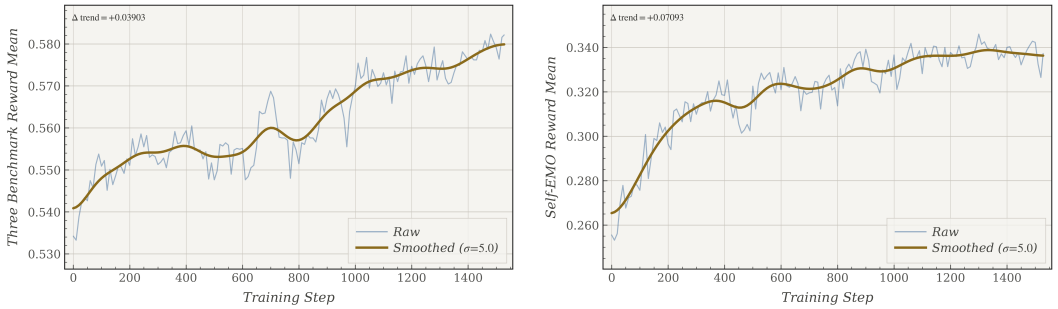



Figure 4: The twelve largest generated persona clusters. Proportions describe the constructed training data, not a population.




(a) Mean ERC diagnostic reward.

(b) Generated-trajectory IoU reward.

Figure 5: Reward trajectories on the held-out training diagnostic set. The raw curves and their smoothed trends are shown without cross-seed error bands.

## C Multi-Task Training Metrics and Analysis

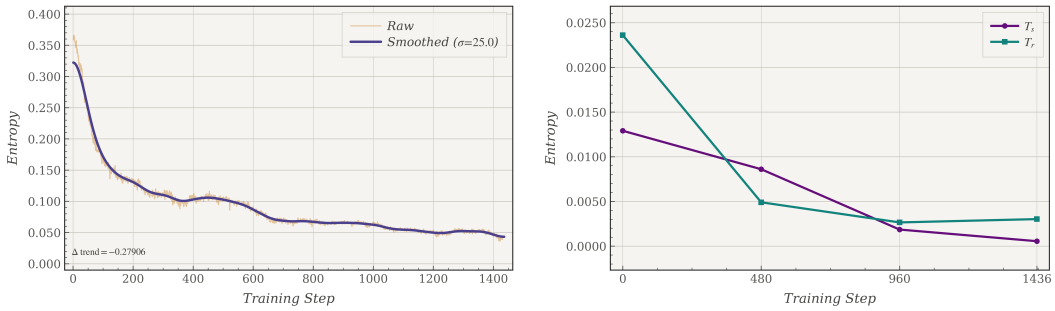
This section reports reward and entropy measurements across training checkpoints. All diagnostic evaluations use the held-out subset derived from the training data rather than the official benchmark test sets. The curves describe the observed training run and do not establish a deterministic relation among tasks.

### C.1 Reward Trajectories

Figure 5 presents the mean ERC benchmark reward and generated-trajectory weighted-IoU reward. Both raw curves and smoothed trends increase overall during training, and no reward collapse is observed in the displayed trajectory. The weighted-IoU trajectory is smoother in this run.

| Cluster label | Count | Cluster label | Count |
|---|---|---|---|
| emotional directness assertive engagement | 1108 | emotional expressiveness interpersonal playfulness reactive curiosity | 590 |
| humor regulated social | 487 | composed practical direct efficient | 329 |
| emotional reactivity defensive confrontation pragmatic dominance | 307 | emotional reactivity impulsive confrontation | 305 |
| empathetic supportive cooperative practical | 279 | emotional restraint direct communication pragmatic interpersonal | 275 |
| emotional vulnerability interpersonal hesitancy reactive defensiveness | 272 | emotional reactivity confrontational impulsivity | 260 |
| emotionally cautious socially adaptive | 259 | humor driven social casual reactive light | 253 |
| direct emotional pragmatic | 210 | calm cooperative pragmatic | 210 |
| emotionally regulated socially engaged assertive receptive | 201 | humor driven social reactivity | 194 |
| expressive playful adaptable supportive | 193 | emotional neutrality direct factuality | 189 |
| reactive defensive unfiltered | 181 | assertive emotionally direct autonomous | 174 |
| emotionally reactive assertive confrontational | 171 | emotional reactivity interpersonal hesitancy anxiety driven | 166 |
| cautious empathetic guarded | 149 | emotional mixed interpersonal casual pragmatic hesitant | 137 |
| procedural calm solution oriented | 133 | emotional expressivity interpersonal reactivity impulsive coping | 129 |
| anxious evasive overthinker | 129 | emotional vulnerability reactive interpersonal seek withdraw defensive sensitivity | 127 |
| pragmatic calm direct supportive | 123 | emotionally restrained pragmatic supportive | 109 |
| reserved practical | 104 | emotional restraint passive compliance | 103 |
| reactive urgent assertive pragmatic | 102 | direct pragmatic emotional control | 102 |
| trauma bound guilt moral compulsion | 100 | emotional reactivity assertive adaptability | 98 |
| pragmatic defensiveness | 93 | emotional reactivity interpersonal evasiveness | 92 |
| humor driven socially adaptive playful critique | 90 | rational calm procedural neutral | 86 |

Table 5: Full generated persona-cluster distribution (part 1).




(a) Policy token entropy.

(b) Field-level entropy for $T_s$ and $T_r$.

Figure 6: Entropy trajectories on the held-out training diagnostic set. All reported final values are below their initial values, although reply-token entropy increases slightly near the end.

### C.2 Joint Entropy Reduction

To evaluate Eq. 5, we track average token-level entropy over training. Policy entropy is measured throughout the run. At the selected checkpoints shown in Figure 6, we separately average token entropy over the response-emotion field $T_s$ and reply field $T_r$ on the same held-out training diagnostic set. Policy and response-emotion entropy decline overall. Reply-token entropy drops early and then increases slightly near the end, and its final value remains below its initial value. These measurements support joint entropy reduction as an empirical pattern.

| Cluster label | Count | Cluster label | Count |
|---|---|---|---|
| emotional reactivity spontaneous expression direct engagement curiosity driven | 84 | emotional reactivity confrontational impulsive | 83 |
| calm restraint polite cooperative rule oriented | 81 | emotional expressiveness persistent engagement nature reactivity idealistic connection | 80 |
| emotionally expressive socially playful | 78 | emotional reactivity defensive defiance persistent worthlessness validation struggle | 78 |
| emotional vulnerability reactive grief empathetic support resilient reframing | 76 | nostalgic guarded withdrawal | 76 |
| calm direct practical | 76 | empathetic supportive pragmatic calm | 74 |
| high energy expressive cooperative optimistic | 72 | desperate urgency hope clinging plea driven defensive irrationality | 70 |
| emotional reactivity defensive pessimism | 68 | emotional fragility grief reactivity | 67 |
| empathetic pragmatic supportive patient | 66 | calm procedural assertive | 65 |
| emotionally restrained casual deescalator | 62 | empathetic supportive practical | 62 |
| emotionally expressive playfully spontaneous socially engaged | 60 | empathetic listener | 59 |
| pragmatic detached skeptical | 59 | pragmatic emotionally restrained direct task oriented | 59 |
| emotional volatility reactive sensitivity | 59 | warmth driven expressive reactive | 58 |
| emotional reactivity assertive defensiveness interpersonal fluctuation | 58 | emotional reactivity expressive friendliness pragmatic cooperativeness | 58 |
| emotional reactivity impulsive seeking connection conflict prone | 57 | calm cooperative professional | 55 |
| pragmatic direct | 55 | socially oriented curious engaging | 52 |
| humor driven social engager | 51 | rational casual defensive fact oriented | 51 |
| social playfulness | 50 | curious neutral direct inquisitive | 50 |

Table 6: Full generated persona-cluster distribution (part 2).

| Model | EDAE | EEIPRC | HRS | CPDE | ERDCPD | Avg. |
|---|---|---|---|---|---|---|
| Seed-1.8 | 14.21 | 12.81 | 15.75 | 13.06 | 13.03 | 13.77 |
| DeepSeek-V3.2 | 9.21 | 10.89 | 10.21 | 10.25 | 9.64 | 10.04 |
| SELF-EMO-8B | 56.32 | 58.83 | 61.98 | 57.09 | 58.16 | 58.48 |

Table 7: Reward percentages stratified by the five largest generated clusters. These are diagnostic reward values, not personality distributions.

# D Response Comparison Before and After Training

## D.1 LLM-as-a-Judge Protocol

The auxiliary-output evaluation uses 500 randomly sampled dialogue examples. For each example, Seed-1.8 and DeepSeek-V3.2 receive the role description, dialogue context, and two outputs identified only as `answer1` and `answer2`. The checkpoint identities are not disclosed. Each judge examines the role description and dialogue context, assesses both answers under the same criteria, and returns the answer that better matches the role and context. The aggregate preferences are reported in Figure 3. Appendix E reports the corresponding human evaluation.

The following examples reproduce the observable structured fields used in this comparison, in Tables 8 and 9. Raw generated chain-of-thought and the judges' long-form rationales are omitted.

## D.2 Example 1: Monica Responds to Joey

## D.3 Example 2: Chandler Reassures Monica

These cases illustrate the criteria used by the judges. The reported preference numbers come from the full 500-example comparison.

**Response Comparison Example 1**

**Role.** Monica is an expressive and socially proactive person, likely in her 20s–30s, who uses casual, engaging questions to foster group connection and share emotional observations.
**Available labels.** neutral, surprise, fear, sadness, joy, disgust, anger.
**Dialogue context.**
Monica: He is so cute. So, where did you guys grow up?
Angela: Brooklyn Heights. Bob: Cleveland.
Monica: How, how did that happen?
Joey: Oh my god. Monica: What?
Joey: I suddenly had the feeling that I was falling. But I'm not.

| | Later checkpoint | Early checkpoint |
|---|---|---|
| Interlocutor emotion | neutral: 0.8; surprise: 0.6 | surprise: 0.95; fear: 0.75; disgust: 0.6 |
| Response emotion | joy: 0.6; surprise: 0.5 | joy: 0.7; neutral: 0.5 |
| Reply | "Yeah, that sounds intense—" | "Wisely, mark 'falling' as a pseudonym for your favor." |

**Judge outcome.** The later-checkpoint output is preferred because it is more conversational and better aligned with the supplied role and context. Its truncated reply also shows that model preference does not imply a flawless response.

Table 8: Observable outputs for the first early-versus-later checkpoint comparison. The source target contains neutral, surprise, and fear with weights 0.85, 0.45, and 0.25.

# E Additional Human and Stability Evaluations

## E.1 Human-Evaluation Protocol

We sample 100 test examples from each dataset. For both Qwen3-8B and SELF-EMO-8B, we retain the first generated analysis for each example. Human annotators judge whether the analysis is plausible given the dialogue. The reference label is provided as supporting information rather than an absolute criterion. The resulting evaluations of 600 model outputs produce the scores in Table 10, where SELF-EMO gains 10 points overall. Across both models, 168 analyses are judged plausible despite disagreeing with the reference label. This result illustrates ambiguity in single-label ERC.

The pairwise comparison uses the same 300 examples. SELF-EMO is preferred in 41 cases and Qwen3 in 11; the remaining comparisons are ties. The non-tied win rate is therefore 78.85%. This study evaluates generated analyses. It does not directly evaluate replay pseudo-labels or generated replies.

## E.2 Rollout Stability

For every example in the three official test sets, we sample eight outputs. First Acc. evaluates the first output, and Majority Acc. uses majority voting. Oracle@8 counts an example as correct when any rollout predicts the reference label. All-8 Correct requires every rollout to be correct. Consistency@8 requires all eight predictions to agree. Overall scores pool all test examples and are therefore sample-weighted. As shown in Table 11, SELF-EMO raises overall First Acc. from 56.11% to 60.13%, All-8 Correct from 39.58% to 51.81%, and Consistency@8 from 59.34% to 78.51%, while Oracle@8 falls from 69.81% to 67.45%.

## E.3 Representative Error Cases

Table 12 gives three cases from the human study. In each case Qwen3-8B is implausible and incorrect, while SELF-EMO-8B is plausible and matches the reference. These examples illustrate common errors.

**Response Comparison Example 2**

**Role.** Chandler is a rational and supportive person, likely in his 20s–30s, who uses structured, future-focused reassurance and gentle persuasion to defuse tension and maintain cooperation.
**Available labels.** neutral, surprise, fear, sadness, joy, disgust, anger.
**Dialogue context.**
Monica: I can't believe her, you know it's just it's so typical.
Chandler: There is going to be a wedding, you are going to throw the bouquet, and...
Monica: Paris?
Chandler: We will take a moonlit walk on the Rue de la.
Monica: Keep talking.

| | Later checkpoint | Early checkpoint |
|---|---|---|
| Interlocutor emotion | anger: 0.6; sadness: 0.4 | neutral: 0.8; surprise: 0.3 |
| Response emotion | joy: 0.3; neutral: 0.7 | neutral: 0.7; joy: 0.4 |
| Reply | "Monica, I get it. But let's not get stuck in this moment. We've got a wedding to plan, a bouquet to toss, and a beautiful evening ahead. Paris is waiting, and I'd rather we focus on the magic of it all than the drama of it not being perfect. What do you think?" | "Monica, you're not going to believe this, but I think you and I both know that whatever happens next, you're going to be the best person to throw the bouquet. Now, let's not waste time pretending like this isn't going to work. We'll make it work. You, me, and a few other people who happen to be really good at making things happen. Now, let's get you dressed and get you to the wedding. You've got a lot to do, and you're not going to do it standing here like this." |

**Judge outcome.** The later-checkpoint output is preferred because its empathetic, collaborative, and future-oriented response better matches the supplied Chandler role. The early output is more directive and less consistent with gentle persuasion.

Table 9: Observable outputs for the second early-versus-later checkpoint comparison. The source target contains disgust and sadness with weights 0.85 and 0.4.

| Model | EmoryNLP | IEMOCAP | MELD | Overall |
|---|---|---|---|---|
| Qwen3-8B | 72.00 | 66.00 | 83.00 | 73.67 |
| SELF-EMO-8B | **79.00** | **76.00** | **96.00** | **83.67** |

Table 10: Human-rated plausibility of generated analyses (%). We sample 100 examples from each test set, so Overall is the unweighted mean of the three columns.

## F Hyperparameter Configuration

We conduct all experiments with the VeRL framework on one node with eight NVIDIA H100 GPUs. The backbone is Qwen3-4B or Qwen3-8B. Table 13 lists the documented training configuration.

The configuration key `trainer.test_freq` controls periodic evaluation on the training-derived diagnostic split; it does not indicate repeated access to the official benchmark test splits.

| Dataset | Model | First Acc. | Majority Acc. | Oracle@8 | All-8 Correct | Consistency@8 |
|---|---|---|---|---|---|---|
| EmoryNLP | Qwen3-8B | 42.89 | 43.39 | **54.47** | 32.11 | 62.60 |
| | SELF-EMO-8B | **43.50** | **43.60** | 48.78 | **38.92** | **79.88** |
| IEMOCAP | Qwen3-8B | 57.18 | 57.61 | **72.09** | 39.37 | 58.23 |
| | SELF-EMO-8B | **60.69** | **61.43** | 70.98 | **49.17** | **73.51** |
| MELD | Qwen3-8B | 60.42 | 61.53 | **74.18** | 42.53 | 58.81 |
| | SELF-EMO-8B | **66.05** | **66.25** | 72.30 | **58.31** | **81.11** |
| Overall | Qwen3-8B | 56.11 | 56.89 | **69.81** | 39.58 | 59.34 |
| | SELF-EMO-8B | **60.13** | **60.48** | 67.45 | **51.81** | **78.51** |

Table 11: Eight-rollout accuracy and stability (%). Overall pools all test examples and is therefore weighted by test-set size. SELF-EMO improves single-rollout accuracy and agreement, while its lower Oracle@8 indicates less sampling diversity.

| Target utterance | Qwen3-8B | SELF-EMO-8B | Ref. |
|---|---|---|---|
| "I don't know if you'd understand." | Predicts *sadness* from implied loneliness and misses the neutral statement of uncertainty. | Uses the surrounding exchange and predicts *neutral*. | neutral |
| "Oh, she was a cruel, cranky, old bitch!... Did Dad say I get the dollhouse?" | Predicts *Mad* from local negative words. | Uses the surrounding excitement about the dollhouse and predicts *Joyful*. | Joyful |
| "Oh absolutely." | Interprets the reply as sarcasm and predicts *angry*. | Connects it to the preceding surprise and excitement and predicts *excited*. | excited |

Table 12: Representative errors from the human-evaluation sample.

| Hyperparameter | Value | Hyperparameter | Value |
|---|---|---|---|
| `train_batch_size` | 64 | `rollout.n` | 8 |
| `max_prompt_length` | 4096 | `rollout.max_num_batched_tokens` | 12400 |
| `max_response_length` | 8192 | `rollout.gpu_memory_utilization` | 0.5 |
| `model.path` | Qwen3-4B/8B | `rollout.tensor_model_parallel_size` | 8 |
| `actor.optim.lr` | $10^{-6}$ | `ref.log_prob_micro_batch_size_per_gpu` | 16 |
| `actor.ppo_mini_batch_size` | 32 | `ref.fsdp_config.param_offload` | True |
| `actor.ppo_micro_batch_size_per_gpu` | 16 | `algorithm.use_kl_in_reward` | False |
| `actor.kl_loss_coef` | 0.001 | `trainer.critic_warmup` | 0 |
| `actor.kl_loss_type` | low_var_kl | `trainer.n_gpus_per_node` | 8 |
| `actor.entropy_coeff` | 0 | `trainer.nnodes` | 1 |
| `actor.fsdp_config.param_offload` | False | `trainer.save_freq` | 480 |
| `actor.fsdp_config.optimizer_offload` | False | `trainer.test_freq` | 10 |
| `rollout.log_prob_micro_batch_size_per_gpu` | 16 | `trainer.total_epochs` | 3 |

Table 13: VeRL hyperparameters used in the experiments.