Adaptive Emotion Management in Human-robot Dialogue using Online Group Relative Policy Optimization

Anna Manaseryan
annamanaseryan@u.boisestate.edu
Computer Science
Boise State University

Casey Kennington
caseykennington boisestate.edu
Computer Science
Boise State University

Abstract

Emotion expression is essential for human-robot interaction, yet current systems rely on static models that cannot adapt to individual users. We present an online reinforcement learning framework that adapts robot emotional behavior policy during live dialogue using binary human feedback. The system integrates a DeBERTa-v3-base emotion classifier and applies Group Relative Policy Optimization (GRPO) in a human-robot dialogue system. At each dialogue turn, the classifier samples a group of emotion candidates and the selected emotion is passed to a generative model that synthesizes a novel robot emotional behavior. We evaluate the system in three experiments: (1) offline supervised fine-tuning followed by GRPO on synthetic dialogue data, (2) a live GRPO training with a human teacher and (3) a final experiment with human participants. Results indicate that the robot was perceived as responsive and emotionally consistent, with high ratings for personality coherence and contextual appropriateness of emotional behaviors. Results further show that online GRPO with human feedback enables effective real-time emotion adaptation in embodied interaction.

1 Introduction

Emotion is central to natural human-robot interaction (HRI), particularly when spoken dialogue is involved. Robots that recognize and convey emotion are perceived as more lifelike, engaging, and socially competent than emotionally neutral robots (Stock-Homburg, 2021; Abdollahi et al., 2023; Fischer et al., 2019). Modeling *which* emotion to display, (e.g., an emotion that signals confusion when the robot needs more information from the user) is as important as *when* to display that emotion. Following the terminology of *dialogue management*, we denote the underlying task of determining when and which emotion to display *emotion management*: at each dialogue turn, a policy must select a contextually appropriate emotion label, which is then realized on the robot as a multimodal behavior that affects and controls robot facial expressions, sounds, drive motions, as well as head and arm movements. Recent work has shown that neural models can *recognize* (McNeill and Kennington, 2019) and large language models (LLMs) can *generate* turn-level emotion in real time (Mishra et al., 2023; Chen and Bian, 2026). Moreover, recent work in reinforcement learning (RL) with scalar or binary human feedback has become a standard recipe for post-training in language modeling (Ouyang et al., 2022; Pandey and Jin, 2026). Together, these advances make it plausible to *adapt* a robot's emotion policy from live user feedback during an ongoing interaction.

Two gaps nevertheless remain. First, existing real-time emotion systems for robots are *static*; the classifier is trained once on annotated corpora and then frozen at deployment, so the model cannot adjust to how a particular user interprets robot emotion (Stock-Homburg, 2021; Mishra et al., 2023). Second, where RL has been applied to social agents or to language-model post-training, the reward signal comes from narrow action spaces (Tamantini et al., 2026; Pereira et al., 2024), simulated users, or offline verifiable signals (Wang et al., 2025a,b) rather than from a real human in a live, interactive, embodied loop.

Closing this gap is challenging because human feedback is inherently noisy and sparse. In particular, applying standard online RL to a classifier under these conditions risks instability: the reward is binary and delayed, the number of feedback events per session is small, and the policy must not drift catastrophically from its initialization. We address these issues with a system built on the Retico incremental processing framework (Michael, 2020; Manaseryan et al., 2025) that performs online Group Relative Policy Optimization (GRPO; Shao et al., 2024) on a compact DeBERTa-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 585–596
August 2–5, 2026.

v3-base emotion classifier fine-tuned with LoRA adapters (Hu et al., 2021). At each turn, the classifier samples a group of candidate emotions; one is selected and produced by the Generate RobotEmotional Displays (GRED) model recently introduced by Baral et al. (2025), a generative model that synthesises a novel multimodal behavior on a Cozmo robot. The participant's yes/no feedback is then propagated to all sampled candidates through a similarity-aware reward, accumulated in a small buffer, and applied asynchronously with a KL penalty to a frozen reference policy that bounds drift. Across 542 live interactions with a single tutor, session-level approval rose from 66.7% in the first session to 94.9% in the final session, compared with 51.4% for the same classifier without online adaptation; a subsequent within-subjects study with ten naïve participants showed that the adapted model is perceived as more coherent in emotion/speech alignment and personality than a non-adapted baseline. Our contributions are as follows:

1. A real-time dialogue system built on Retico (Michael, 2020; Manaseryan et al., 2025) that integrates automatic speech recognition, a child-level language model, an emotion-classifier policy, a generative multimodal behavior model, and robot execution in a single end-to-end interactive loop.

2. An online-GRPO formulation applied to a discrete emotion classifier in a live HRI loop—demonstrating that GRPO's group-relative advantage mechanism is stable under the sparse, noisy binary feedback characteristic of real interaction, without requiring pre-collected preference data.

In the following section we compare our work to others, then explain our methodology. We then explain three experiments where we (1) used synthetic data to bootstrap the emotion manager, then (2) an experiment where a single individual provided reinforcement signals to help the model learn in context, then (3) a final experiment that evaluated the model on multiple human participants. We then offer limitations and conclude.

2 Related Work

Affective robotics research consistently finds that emotionally expressive robots are rated as more lifelike and socially competent than neutral counterparts, especially when displays are contextually appropriate (Stock-Homburg, 2021; Ottoni and Cerqueira, 2024; Fischer et al., 2019; García-Corretjer et al., 2022; Abdollahi et al., 2023).[1] Across these systems, however, the emotion policy is authored by designers or trained once offline and then frozen at deployment, so individual differences in how users interpret robot affect cannot be accommodated after the system is used during inference time.

A parallel line of work addresses the *recognition* of emotion from dialogue, the sub-task our classifier performs at each turn. Emotion Recognition in Conversation (ERC) has converged on transformer-based encoders evaluated on benchmarks such as MELD and IEMOCAP (Poria et al., 2019), with representative systems including DialogueRNN (Majumder et al., 2019) and the RoBERTa-based EmoBERTa (Kim and Vossen, 2021), and more recent LLM-driven predictors of turn-level emotion (Mishra et al., 2023; Chen and Bian, 2026). These models are trained once on annotated corpora and deployed as fixed classifiers. We build on this tradition by starting from a DeBERTa-v3 encoder (He et al., 2021), but treat the classifier as a *policy* that is further optimized online against live binary feedback.

The emotion label we select is produced on the robot by a generative behavior module, linking our work to *expressive behavior generation*. Early systems relied on hand-authored behavior libraries or fixed emotion-to-primitive mappings; recent work replaces these with generative models. Mahadevan et al. (2024) prompt LLMs to produce parameterized control code for expressive motion, and Baral et al. (2025) introduce GRED, a generative–discriminative architecture that synthesizes novel multimodal Cozmo behaviors from an emotion label. We adopt GRED as the downstream generator.

Our central contribution (online policy optimization from real human feedback) sits at the intersection of RL for dialogue, human-robot interaction, and recent preference-based and binary-feedback RL. Earlier work optimized dialogue policies online from real users (Gašić et al., 2013; Su et al., 2016a,b), but used task-completion reward on Gaussian-process or small neural managers rather

[1]The literature often uses the terms *affect* and *emotion* interchangeably. Here, we use *emotion* to differentiate that we are focusing on more cultural-level aspects of emotion signals rather than lower-level affect (Barrett, 2017).

than modern transformer policies. Interactive-RL-from-humans research (TAMER) (Knox and Stone, 2009), COACH and its deep extension (MacGlashan et al., 2017; Arumugam et al., 2019), preference-based RL (Christiano et al., 2023), and PEBBLE (Lee et al., 2021), surveyed by Cruz and Igarashi (2021) has established that scalar or binary human feedback can drive RL, but existing demonstrations are in 'gridworlds' or low-dimensional robot control, not dialogue-driven emotion policies with multimodal embodied output. Recent work has converged on PPO- and GRPO-style optimization (Schulman et al., 2017; Shao et al., 2024; Guo et al., 2025) with binary or verifiable signals (Wang et al., 2025a,b; Pandey and Jin, 2026), but these works optimize text generation offline on static preference or verifier datasets. Finally, RL in HRI has focused on navigation and manipulation (Pereira et al., 2024; Tamantini et al., 2026) rather than emotion. Each subfield therefore addresses one piece in isolation. To our knowledge, no prior system combines (i) online GRPO, (ii) real feedback from a human during live dialogue, and (iii) multimodal embodied behavior from a downstream generative model; our work fills this gap.

3 Method

In this section, we explain our methodology. We first explain two important components in our full system: (1) the model that generates emotional robot behaviors and (2) the model that speaks like a child. We then explain how our emotion management model works.

GRED The emotion label selected by the GRPO policy must be realized as an observable robot behavior. We use GRED (Generate Robot Emotional Displays) (Baral et al., 2025), a generative model built on GPT-2 Medium that maps an emotion label to a novel sequence of parameterized Cozmo actions (facial displays on an OLED screen, vocalizations, driving patterns, head angles, and lift-arm movements). GRED was fine-tuned on 600 hand-designed behavior sequences (35–40 per emotion) across six grouped emotion categories (similar to Torres-Fonsesca and Kennington (2022)): *Anger/Frustration*, *Confusion/Sorrow/Boredom*, *Disgust/Surprise/Fear*, *Interest/Desire*, *Joy/Hope*, and *Understanding/Gratitude*, where each behavior is represented as a sequence of symbolic action tokens that are language-model readable (e.g., `display_oled_face_image-face2 turn_in_place 30 230`). An adversarial training regime incorporating feedback from an emotion classifier (EMRO, also reported in Baral et al. (2025)) and a novelty metric ensures that the generated behaviors are both emotionally accurate and diverse across repeated generated behaviors.[2] At inference time, GRED receives *only the emotion label* from the emotion manager—not the dialogue transcript—and generates a new behavior, which is then executed on the Cozmo robot. The generated behavior is therefore conditioned on the emotion category alone; dialogue content influences behavior only indirectly, through the emotion classifier's choice of label. This design follows the original GRED architecture and is discussed as a limitation in Section 5.

ChildLM The robot's spoken responses are generated by ChildLM (Levandovsky et al., 2025), a compact generative language model trained to produce utterances consistent with emergent and toddler-level speech (ages 1–3), rather than adult-targeted text. The child-like speech style was chosen for two HRI-motivated reasons: the Cozmo robot's small physical form factor and expressive face naturally invite a child-like social persona, and simple, short utterances reduce the cognitive load of parsing robot speech during live interaction, keeping participants' attention on the emotional displays (Plane et al., 2018). ChildLM is trained on transcribed child–caregiver dialogue from CHILDES (Macwhinney, 2000), using a three-stage recipe: (i) pre-training from scratch on filtered CHILDES tokens, (ii) supervised chat fine-tuning with a SmolLM2-style template on high-coherence subsets of caregiver: child turns, and (iii) GRPO with multiple reward signals to balance child-like surface forms with answer coherence. The resulting checkpoint is a ∼155M-parameter model that, in prior work, matched or exceeded larger SmolLM2 variants on automated coherence scoring against caregiver questions drawn from CHILDES. In our system, ChildLM is held fixed across all experimental conditions: only the emotion-classification policy is trained, so differences in perceived interaction quality attributable to language are shared across the Baseline and GRPO models.

[2] Their work reported *novelty* as an important part of the loss function so the robot would not appear to repeat the same emotional behaviors.

Field	Content
Seeded topic	drawing shapes
Seeded style	question
Target emotion	Joy/Hope
Human utterance	"Should we draw a circle?"
Robot response	"Together? Yes! Let's color them!"

Table 1: Example of a generated human–robot interaction used for training (seeded topic, style, and target emotion; generated utterances).

Emotion Management with GRPO We frame emotion selection as a reinforcement learning problem because the reward signal—a live human's binary approval—is not available as labeled training data before deployment. Supervised fine-tuning requires pre-labeled examples of correct emotion choices per dialogue context; GRPO enables the policy to improve from sparse yes/no feedback online, without collecting a per-user labeled corpus. The output space is a discrete emotion class, but the *preference* over that space is user-specific and must be discovered during interaction. The **policy** π_θ is a DeBERTa-v3-base encoder (He et al., 2021) that maps a dialogue turn to a distribution over six emotion classes.

Following GRPO methodology (Shao et al., 2024), we sample a group of G candidate emotions for each state. For each input, a group of G candidate emotions is sampled from the policy. The reward R_i for each candidate is determined by human feedback (or a proxy during bootstrapping). Advantages $A_i = (R_i - \mu_R)/\sigma_R$ are computed by normalizing each reward relative to the group mean μ_R and standard deviation σ_R, so that the policy learns which candidates were better *than the group average* rather than against an absolute reward threshold. The model is then optimized using a clipped surrogate objective with a KL penalty to prevent the policy from drifting too far from a reference model π_{ref}. The KL penalty β bounds how far the updated policy π_θ may drift from the reference policy π_{ref} (the frozen SFT checkpoint), preventing catastrophic forgetting under sparse feedback:

$$L_{\text{total}} = L_{\text{clip}} + \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \qquad (1)$$

where L_{clip} is the mean clipped advantage across the group, and β controls the strength of the trust-region constraint.

System Architecture Figure 1 depicts our full dialogue system, implemented in the Retico incremental processing framework (Michael, 2020; Manaseryan et al., 2025). Speech input is captured from a table microphone and transcribed with Whisper Automatic Speech Recognition (ASR) (Radford et al., 2022). The transcript is passed to ChildLM, which generates the robot's reply at a child-level utterance. Then, the DeBERTa-based emotion classifier selects one of six emotion labels for the current turn; this module is the only component trained with online GRPO from sparse binary feedback, as detailed in Section 4.2. The encoder does not modify or condition the language model's output; ChildLM generates the spoken response independently, and the emotion classifier operates in parallel over the same dialogue context to select an emotion label. The chosen label informs GRED, which synthesizes a multimodal behavior (robot facial display, short vocalization, head and lift motion, and driving pattern) executed on a Cozmo robot. All modules that produce intermediate outputs are logged for later analysis. The behaviors are only for emotion; the robot is only a social robot that did not perform any other task actions. The emotion space does not include a neutral class; the system is designed to always produce an emotional display. This design follows the GRED behavior model, which pre-defines behaviors for exactly the six emotion clusters listed above. "When to display no emotion" is outside the scope of this work and noted as a limitation.

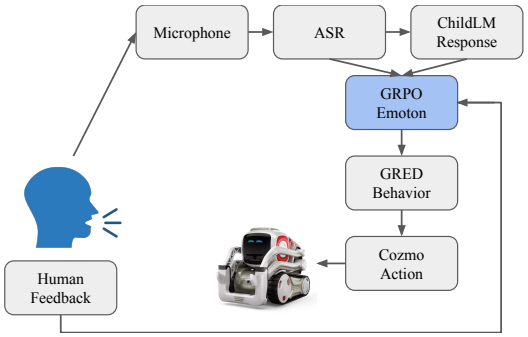

Figure 1: End-to-end pipeline: Microphone, Whisper ASR, ChildLM response generation, GRPO-trained emotion classifier, GRED behaviour synthesis, and Cozmo execution. Binary feedback closes the learning loop for the emotion policy (Experiment 2).

4 Experiments

4.1 Experiment 1: GRPO Bootstrapping with Synthetic Data

Procedure We trained the emotion classifier in two stages: *supervised fine-tuning* (SFT) fol-

lowed by GRPO on a purpose-built synthetic dataset, since no labeled corpus of child-directed human–robot dialogues with emotion annotations was available. The dataset was generated with NVIDIA NeMo Data Designer (The NeMo Data Designer Team, 2025) using an `LLMTextColumnConfig` that injects three variables into a Jinja-templated prompt; one of six target emotion clusters, a conversation topic (e.g., *playing with toys*, *learning numbers*), and an interaction style (e.g., *question*, *game*). The prompt constrains each generation to at most ten words, enforcing a toddler-like persona and a fixed `human: . . . cozmo: . . .` dialogue schema. We generated 10,000 unique human–robot interactions. The emotion label space is the six emotion-cluster grouping of Baral et al. (2025) (Anger/Frustration, Confusion/Sorrow/Boredom, Disgust/Surprise/Fear, Interest/Desire, Joy/Hope, and Understanding/Gratitude) chosen because GRED provides pre-trained multimodal Cozmo behaviors for exactly these clusters, so the upstream classifier can drive the downstream behavior generator without relabeling. Table 1 shows one generated interaction.

We first fine-tuned (SFT) DeBERTa-v3-base on a 10% subset of this dataset (∼1 000 samples), with the remaining 90% reserved for GRPO training. Each input was formatted as `HUMAN: {utterance} ROBOT: {response}` and passed through the encoder; mean-pooled hidden states (excluding padding) were projected by a classification head (dropout p=0.3 + linear layer) trained with cross-entropy loss (AdamW, lr = 2e-5, batch 16, 5 epochs). This SFT checkpoint is used in two downstream roles: it initializes both the trainable and the reference policy for the Experiment 1 GRPO run, and it serves as the *Baseline* (non-adapted) model compared against the adapted GRPO policy in Experiment 3.

GRPO training then loaded this SFT model as both the trainable policy π_θ and a frozen reference policy π_{ref}. We applied LoRA adapters to the top two encoder layers and the classification head; all other parameters remained frozen. The reward combined four terms: label agreement (α=1.0), confidence margin (β=0.02), dialogue consistency (γ=0.3), and entropy regularization (δ=0.6). The policy was optimized with AdamW (lr = 5e-5, batch 32, KL coefficient λ=0.4, group size G=10) for up to 5,000 steps.

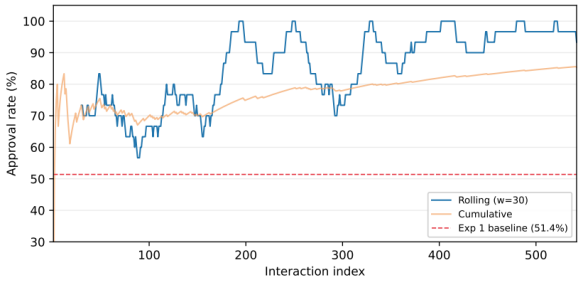

Figure 2: Rolling-window approval rate (window = 30) over 542 chronologically ordered interactions. Dashed line: Experiment 1 baseline (51.4%). A clear shift occurs after approximately 150 interactions; the rolling rate remains above 90% thereafter, apart from one transient dip around interactions 251–300.

Metrics & Baselines The model was evaluated on two levels. First, **offline performance** was measured as accuracy and macro-averaged F1 on a held-out synthetic test set (∼900 samples drawn from the 90% GRPO split). Second, the trained model was deployed in the full Retico pipeline with a Cozmo robot for a **live evaluation** with one participant (70 dialogue turns total). After each turn, the participant provided binary yes/no feedback indicating whether the robot's emotional display was appropriate; the **approval rate** (fraction of YES responses) served as the primary human-facing metric. A post-study questionnaire collected Godspeed-style impressions and emotion-specific Likert items (Agree/Neutral/Disagree).

Results On a held-out synthetic test set, the GRPO-trained model achieved 89.8% accuracy and a macro-averaged F1 of 0.898 across all six emotion classes. The mean KL divergence from the reference policy was 0.032, confirming that the policy remained close to the SFT initialization. Analyses revealed two systematic error patterns: Interest/Desire was frequently misclassified as Joy/Hope, and confusion was common between Disgust/Surprise/Fear and Anger/Frustration.

However, when deployed in live interaction, the model achieved an approval rate of only **51.4%** (36 YES out of 70 turns). This is well above a random baseline, but not accurate enough for use in a live, interactive system. The questionnaire indicated that the participant found the robot responsive and kind, but rated emotional timing as inappropriate (Disagree) and emotional match to speech as inconsistent (Neutral). In open-ended feedback, the participant noted: *"Sometimes it wasn't as excited. It wouldn't move as much or twirl around. . . it would*

happen in situations that didn't fit."

These results demonstrate a gap between synthetic and real-world performance: while the model learned accurate emotion boundaries from synthetic labeled data as a means to bootstrap the model, it failed to consistently produce contextually appropriate emotions during live dialogue. This motivated Experiment 2, in which the policy is adapted online from real human feedback.

4.2 Experiment 2: Interacting with a Human Tutor

Procedure We used Retico (Michael, 2020; Manaseryan et al., 2025) for our interaction pipeline.[3]

One adult participant (an undergraduate computer science student with no prior experience with the robot) interacted with the Cozmo robot across 18 sessions (approximately 9 hours in total), organized in 30-minute blocks. The participant was instructed to ask short questions typical of conversations with a 2–3-year-old child (e.g., likes and dislikes, feelings, everyday activities, imaginary scenarios) and to explore diverse conversation topics. After each dialogue turn, the participant provided binary yes/no feedback using a keyboard indicating whether the robot's emotional display was appropriate. Full participant instructions are provided in Appendix A.

At each turn, the policy sampled a group of $G{=}4$ candidate emotion labels; one was selected at random and displayed via GRED on the robot. The participant's binary feedback was propagated to all four candidates through a similarity-aware reward function: if the participant said YES, the displayed emotion received $R{=}3$, semantically similar emotions (e.g., Interest↔Joy) received $R{=}2$, and dissimilar emotions received $R{=}0$. For NO feedback, the rewards were reversed ($R{=}0$, 1, 2 respectively). Only LoRA adapters were updated. Asynchronous locking ensured inference and training never overlapped on the same weights. Adapter weights and optimizer state were checkpointed after each batch update and restored at the start of each subsequent session, so learning persisted across the 18 sessions while the reference policy remained fixed.

Metrics & Baselines The primary metric is the **approval rate**: the proportion of turns in which the participant judged the robot's emotional display as appropriate (YES). All 542 interactions were ordered chronologically and analyzed in three ways:

1. as a **binned** into consecutive groups of 50 interactions,
2. a **cumulative** running proportion, and
3. **rolling window** of 30 interactions to smooth local fluctuations

We additionally report **session-level** approval (the fraction of YES responses within each 30-minute tutoring block) so that the adaptation trajectory can be summarized at its start and endpoint. Per-emotion approval was computed across all interactions. Training stability was tracked via **KL divergence** between the active LoRA-adapted policy and the frozen reference policy, recorded at each of the 61 batch updates.

The **offline-GRPO baseline** from Experiment 1 (51.4% approval across 70 turns with a human participant; Section 4.1) offers a baseline for these numbers: it is the same classifier without any live adaptation.

Interactions	n	YES	NO	YES %
1–50	50	37	13	74.0
51–100	50	33	17	66.0
101–150	50	35	15	70.0
151–200	50	45	5	90.0
201–250	50	47	3	94.0
251–300	50	37	13	74.0
301–350	50	46	4	92.0
351–400	50	48	2	96.0
401–450	50	46	4	92.0
451–500	50	49	1	98.0
501–542	42	40	2	95.2
Total	542	463	79	85.4

Table 2: Approval rate in consecutive bins of 50 interactions each, in chronological order across all 18 sessions. This is complementary to session-level approval (first session 66.7%, final session 94.9%), defined in Section 4.2.

Results Over 542 interactions across 18 sessions (approximately 9 hours), the participant provided 463 YES and 79 NO responses, yielding an overall approval rate of 85.4%. This overall figure averages the full trajectory; session-level approval (defined above) moves from 66.7% in the first session to 94.9% in the final session, and the offline-GRPO baseline without live adaptation reached only 51.4% (Section 4.1). This model constitutes our trained emotion manager.

[3] Specifically, we required two separate CUDA-ready machines to run our full system split across two hosts connected via ZeroMQ (Akgul, 2013).

Table 3: Per-emotion approval rate across all 542 interactions.

Emotion class	YES	Total	YES %
Anger/Frustration	46	50	92.0
Understanding/Gratitude	95	108	88.0
Interest/Desire	105	121	86.8
Joy/Hope	86	101	85.1
Disgust/Surprise/Fear	30	36	83.3
Confusion/Sorrow/Boredom	101	126	80.2
Total	463	542	85.4

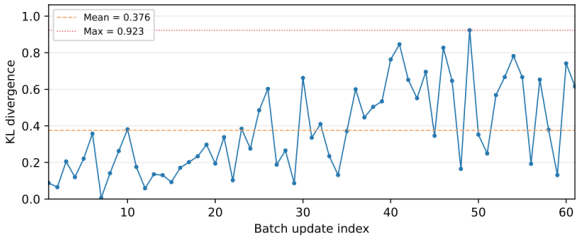

Figure 3: KL divergence between the active policy and the frozen reference policy over 61 batch updates. The penalty ($\beta = 0.04$) kept divergence bounded (max = 0.923), preventing unbounded drift from the reference policy.

Approval trends. Table 2 and Figure 2 present the approval trajectory. During the first 150 interactions, the binned rate fluctuated between 66% and 74%, reflecting the policy's initial exploration phase. A clear shift occurred around interaction 150: the rate rose to 90% and remained above 92% for the remaining 342 interactions, with individual bins reaching 96–98%. One exception is the bin covering interactions 251–300, where the rate temporarily dropped to 74% before recovering; this corresponded to a block in which the participant explored unusually short or ambiguous prompts.

Per-emotion analysis. Table 3 reports the approval rate by emotion class. All six classes exceeded 80%. The emotion distribution was unbalanced: four classes (Confusion, Interest, Understanding, Joy) accounted for 84% of displayed emotions, reflecting the natural topic distribution in conversation.

Training stability. The policy performed 61 batch updates over 542 interactions (Figure 3). The mean KL divergence was 0.376, with a maximum of 0.923; the penalty term ($\beta = 0.04$) successfully prevented unbounded drift from the reference policy. Total loss remained stable (mean = 0.149) with no upward trend. The mean clip fraction was 0.743, reflecting the online setting in which importance ratios frequently exceeded the clipping threshold; despite this, the policy continued to improve.

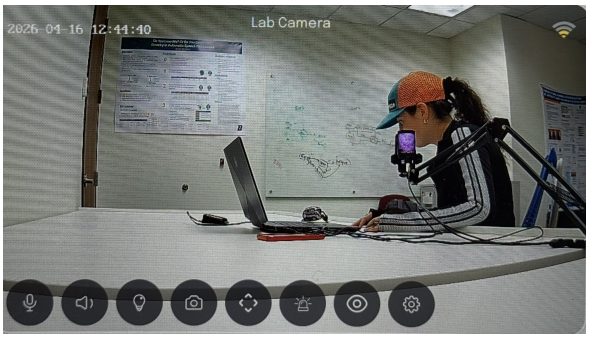

Figure 4: A participant interacting with the Cozmo robot during Experiment 3. The robot is placed on a table at a natural conversational distance; the participant speaks freely while Cozmo responds with ChildLM-generated speech and GRED multimodal emotional behaviors.

Questionnaire trends. The participant completed seven questionnaires at regular intervals during the 18 sessions. A clear progression emerged: *"emotional expressions appeared at the right time"* shifted from Neutral to consistent Agree by the midpoint of the study, while all other emotion items (engaging, understood-feelings, important, comfortable) were rated Agree throughout. In open-ended responses, the participant's language evolved from noting occasional randomness (*"movements are a little disorganised"*) to recognising a coherent personality (*"I can see his personality a lot more—I can tell when something makes him shy vs happy"*).

4.3 Experiment 3: Evaluation with Human Participants

Procedure To assess whether the GRPO-adapted emotion manager generalizes beyond the single tutor of Experiment 2, we conducted a within-subjects study with ten participants (see Figure 4 for the experimental setup). Each participant interacted with the robot twice, once with the *Baseline* (SFT-only) model from Experiment 1 and once with the *GRPO* model, for approximately 20 minutes per session, without knowing which model they were interacting with. The interaction order was counterbalanced: half of the participants began with the Baseline and half began with the GRPO model. To control for content effects, the questions a participant asked during the first session were printed and provided for the second session, so both models received the same conversational input from each participant. After each session, participants completed a questionnaire; after both sessions, they received a $12 Amazon gift card.

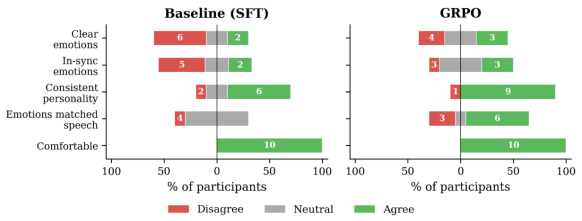

Figure 5: Selected questionnaire responses for the Baseline (SFT) and GRPO models ($N = 10$). *"Emotions matched speech"* shows the largest difference between models; *Comfortable* is identical, confirming that the difference is specific to the emotional dimension.

Metrics & Baselines Participants rated the robot on 16 Godspeed-style items on a 5-point scale (e.g., *Confusing emotions—Clear emotions*) and seven emotion-specific Likert items (Agree / Neutral / Disagree) targeting the contextual appropriateness, timing, and engagement value of the robot's emotional displays. A final open-ended item invited participants to describe any emotions that felt random or out of place. Tables 4 and 5 show the items that differentiate the two models; complementary items are summarized in the text. We compare the SFT model from Experiment 1 as the baseline with the final model improved from live interactions with the human tutor using GRPO from Experiment 2. We note that this comparison conflates two effects: the gap between synthetic training data and real dialogue, and the effect of personalized online adaptation. To isolate personalization, one would need a third condition: a model trained on real (non-personalized) dialogue data without online adaptation (which we did not collect).

Results The clearest difference between models appeared on the item *"The robot's emotions matched what it was saying"*: six of ten participants agreed for the GRPO model, compared with zero for the Baseline, where six were neutral and four disagreed (Figure 5, Table 5). This item most directly targets the objective of the emotion policy—selecting an emotion that fits the conversational context.

The Godspeed-style items show a consistent, though modest, trend in the same direction (Table 4). *Consistent personality* showed the largest gap (GRPO = 4.50 vs. Baseline = 3.70): nine of ten participants rated GRPO at 4 or above, compared with six for the Baseline. *In-sync emotions* (3.00 vs. 2.40) and *Clear emotions* (2.70 vs. 2.20) both trend toward GRPO, though the differences are small on a 5-point scale with ten participants and should be interpreted with caution. *"Helped me understand how the robot felt"* shifted by only one participant toward Agree for GRPO (6 vs. 5), while both models received unanimous Agree on comfort (Table 5), confirming that the observed differences are specific to the emotional dimension.

Table 4: Selected Godspeed-style items (1–5 scale; higher = more positive). Means over $N = 10$ participants.

Item	Baseline	GRPO
Confusing / Clear emotions	2.20	2.70
Out-of-sync / In-sync emotions	2.40	3.00
Inconsistent / Consistent personality	3.70	4.50

Table 5: Emotion-specific Likert items (A = Agree, N = Neutral, D = Disagree; $N = 10$).

	Baseline			**GRPO**		
Item	A	N	D	A	N	D
Emotions matched speech	0	6	4	6	1	3
Helped understand feelings	5	1	4	6	1	3
Comfortable	10	0	0	10	0	0

Open-ended responses provide additional qualitative context. One participant described the Baseline's behavior as "wide-eyed or scared . . . which seemed very strange," but characterized the same robot under GRPO as "less confused and more able to state his opinions," adding that occasional face-hiding "just made him seem like he enjoyed hiding." Another participant noted that under GRPO the robot "did a lot better this time . . . it felt like I was talking to a child," though this participant interacted with the Baseline first, so the improvement may partly reflect a warm-up effect. Several participants reported that the robot's expressions occasionally felt random under both conditions, suggesting that the behavior generation system itself—not only the emotion policy—limits the perceptual clarity of the displays.

In summary, the GRPO model's clearest advantage over the Baseline is a substantially better perceived match between emotions and speech content which was the core objective of the emotion policy. Personality consistency also improved noticeably. Both models were rated identically on comfort, which is also a positive result as the participants never felt uncomfortable. These results suggest that online GRPO training primarily improves emotion–context alignment, while its effect

on broader interaction quality is constrained by factors outside the emotion policy's control.

5 Limitations

Several factors limit the conclusions from Experiment 3. First, we noted minor **order and novelty effects**; despite balancing the order, participants tended to favor the second interaction, likely due to a "warm-up" period or an expectation that the second model would be better. Second, **differences in participant backgrounds** from total beginners to those familiar with earlier, non-emotional versions of the robot, led to a wide range of ratings. Third, **unclear behaviors** in the GRED motor commands caused confusion; the same expression was sometimes seen as "angry" by one person but "happy" by another. This was made worse by **meaningless childish responses** from ChildLM. Participants often rated the emotion as poor if the robot's speech did not make sense, even if the physical display was actually correct for the situation. A further limitation of Experiment 3 is that the baseline (Experiment 1 SFT model) was trained on synthetic data, making it impossible to fully disentangle the effect of domain shift (synthetic to real dialogue) from the effect of personalized online adaptation. Future work should include an intermediate condition trained on real but non-personalized data. Additionally, ASR errors from Whisper may have introduced noise into the emotion classifier's input; we did not separately measure the impact of transcription errors on emotion selection quality. Finally, with only $N = 10$, this small study needs to be confirmed with a larger group, but the results are in favor of the GRPO emotion manager model.

Experiments 1 and 2 relied on a single tutor, whose personal interpretation of appropriate robot emotion shaped the learned policy. The adapted model in Experiment 3 was therefore evaluated for generalization to new users, but cannot be assumed to generalize to all individuals or cultural contexts. Single-user tutoring is an acknowledged limitation of our online-adaptation paradigm; multi-user tutoring or preference aggregation are directions for future work.

Additionally, the binary keyboard feedback mechanism is a laboratory proxy; in real HRI deployments, users cannot be expected to type responses during interaction. More naturalistic feedback signals—such as facial expressions, head nods, or explicit verbal feedback—would be needed for practical deployment, and the adaptation dynamics reported here may not transfer to those modalities. Regarding adaptation speed, our data suggest that approximately 150 interactions are needed before the rolling approval rate consistently exceeds 90%; this represents a practical constraint on how quickly the system can be tuned to a new user.

Because GRED is conditioned only on the emotion label and not on dialogue content, two turns with the same predicted emotion class may produce identical or near-identical behaviors regardless of conversational context. Coupling the behavior generator more tightly to dialogue content is a direction for future work.

6 Conclusion

In this paper, we trained and evaluated an emotion manager which is an online reinforcement-learning framework for adapting a robot's emotion policy from real human feedback during live dialogue. We embedded a compact transformer emotion classifier inside an incremental dialogue pipeline, fine-tuned it with GRPO on a purpose-built synthetic dataset, and then continued to update it during interaction using binary yes/no responses from a human. Our three experiments showed the policy progression: an initial offline-trained model behaved well in synthetic tests but poorly in live interaction, online tutoring produced a sustained improvement, and the subject evaluation with participants showed that the adapted model is perceived as more coherent in emotion–speech alignment and personality consistency than the baseline. Together, the results suggest that online GRPO with human feedback is a viable mechanism for adapting emotional behavior in real time.

For future work, we plan to develop more discriminable behavior primitives to ensure the emotion policy's choices are more readable for participants. Furthermore, we aim to expand the reinforcement learning state beyond speech context by incorporating multi-modal inputs, such as **vision** and **internal robot states**, while exploring a tighter coupling between the emotion policy and the language model to allow for joint adaptation.

Acknowledgements

We want to thank the anonymous reviewers for their helpful feedback and comments. Thanks to Tenneh Kanneh and other members of the SLIM Group

for their help with experiments. This material is based upon work supported by the National Science Foundation under Grant No. 2140642.

References

Hojjat Abdollahi, Mohammad H. Mahoor, Rohola Zandie, Jarid Siewierski, and Sara H. Qualls. 2023. Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. *IEEE Transactions on Affective Computing*, 14(3):2020–2032.

Faruk Akgul. 2013. *ZeroMQ*. Packt Publishing.

Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L. Littman. 2019. Deep reinforcement learning from policy-dependent human feedback. *Preprint*, arXiv:1902.04257.

Rista Baral, Bethany Grenz, and Casey Kennington. 2025. Recognizing and generating novel emotional behaviors on two robotic platforms. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 21503–21510.

Lisa Feldman Barrett. 2017. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.*, 12(1):1–23.

Yanrong Chen and Xihan Bian. 2026. Real-time synchronized interaction framework for emotion-aware humanoid robots. *Preprint*, arXiv:2601.17287.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

Christian Arzate Cruz and Takeo Igarashi. 2021. A survey on interactive reinforcement learning: Design principles and open challenges. *Preprint*, arXiv:2105.12949.

Kerstin Fischer, Malte Jung, Lars Christian Jensen, and Maria Vanessa aus der Wieschen. 2019. Emotion expression in hri – when and why. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 29–38.

Marialejandra García-Corretjer, Raquel Ros, and Roger Mallol. 2022. Empathy as an engaging strategy in social robotics: a pilot study. *User Modeling and User-Adapted Interaction*, 33:1–39.

Milica Gašić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013. On-line policy optimisation of bayesian spoken dialogue systems via human interaction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8367–8371. IEEE.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

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

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *Preprint*, arXiv:2108.12009.

W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16.

Kimin Lee, Laura Smith, and Pieter Abbeel. 2021. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *Preprint*, arXiv:2106.05091.

Enoch Levandovsky, Anna Manaseryan, and Casey Kennington. 2025. Learning to speak like a child: Reinforcing and evaluating a child-level generative language model. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 370–382.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*, pages 2285–2294. PMLR.

Brian Macwhinney. 2000. The childes project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8.

Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative expressive robot behaviors using large language models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 482–491.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Anna Manaseryan, Porter Rigby, Brooke Matthews, Catherine Henry, Josue Torres-Fonseca, Ryan Whetten, Enoch Levandovsky, and Casey Kennington. 2025. rrSDS 2.0: Incremental, modular, distributed, multimodal spoken dialogue with robotic platforms. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Avignon, France. Association for Computational Linguistics.

David McNeill and Casey Kennington. 2019. Predicting human interpretations of affect and valence in a social robot. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany.

Thilo Michael. 2020. Retico: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 49–52, 1st virtual meeting. Association for Computational Linguistics.

Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. 2023. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI*, Volume 10 - 2023.

Lara Toledo Cordeiro Ottoni and Jés de Jesus Fiais Cerqueira. 2024. A systematic review of human–robot interaction: the use of emotions and the evaluation of their performance. *International Journal of Social Robotics*, 16(11):2169–2188.

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 Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Punya Syon Pandey and Zhijing Jin. 2026. Binaryppo: Efficient policy optimization for binary classification. *Preprint*, arXiv:2602.02708.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2024. Deep emotion recognition in textual conversations: A survey. *Preprint*, arXiv:2211.09172.

Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. 2018. Predicting perceived age: Both language ability and appearance are important. In *Proceedings of SigDial*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *Preprint*, arXiv:1905.02947.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Ruth Stock-Homburg. 2021. Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research. *International Journal of Social Robotics*, 14.

Pei-Hao Su, Milica Gasic, Nikola Mrkšić, Lina M Rojas Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016a. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2431–2441.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016b. Continuously learning neural dialogue management. *Preprint*, arXiv:1606.02689.

Christian Tamantini, Alessandro Umbrico, Francesca Fracasso, Gloria Beraldo, Gabriella Cortellessa, and Andrea Orlandini. 2026. Adapting virtual agent interaction style with reinforcement learning to enhance affective engagement. *Frontiers in Digital Health*, Volume 7 - 2025.

NVIDIA The NeMo Data Designer Team. 2025. Nemo data designer: A framework for generating synthetic data from scratch or based on your own seed data. `https://github.com/NVIDIA-NeMo/DataDesigner`. GitHub Repository.

Josue Torres-Fonsesca and Casey Kennington. 2022. HADREB: Human appraisals and (English) descriptions of robot emotional behaviors. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, Yuan Li, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025a. Rlver: Reinforcement learning with verifiable emotion rewards for empathetic agents. *Preprint*, arXiv:2507.03112.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Ellie Evans, Daniel Egert, Hoo-Chang Shin, Felipe Soares, Yi Dong, and Oleksii Kuchaiev. 2025b. Rlbff: Binary flexible feedback to bridge between human feedback & verifiable rewards. *Preprint*, arXiv:2509.21319.

A Participant Instructions

A.1 Experiment 2: Human Tutor

About the robot. The robot can move, show facial expressions, make sounds, and speak. It responds in a simple way, similar to a young child. It cannot see and does not retain memory of previous answers.

Task. Ask the robot a variety of short questions you might ask a 2–3-year-old child, such as questions about likes and dislikes, feelings, everyday activities, imaginary scenarios, or its thoughts on different topics. Explore different types of conversations rather than staying on one theme, and try to ask as many questions as possible within the allotted time.

Feedback. After each question, wait for the robot to respond. You will then be prompted on screen to type whether the emotion was appropriate (`yes` or `no`).

A.2 Experiment 3: Human Evaluation

About the robot. Same as Experiment 2.

Task. Same as Experiment 2, except that the feedback is used for evaluation only and does not update the model. Each participant interacts with two versions of the robot (order counterbalanced) and fills out a questionnaire after each session.