Making Human-Robot Interaction Emotionally Appropriate with Real-Time Reinforcement Learning

Anna Manaseryan
Boise State University

Casey Kennington
Boise State University

SIGdial

INTRODUCTION

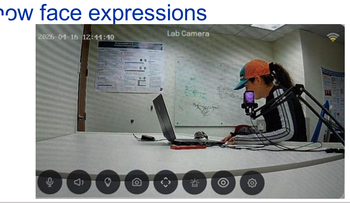
Problem: Robots interacting with humans need to express appropriate, context-aware emotions. Training models on static data fails to capture the dynamic nature of real human-robot interaction.

Solution: We train an emotion recognition model using a two-phase pipeline, first on synthetic data for initialization, then refined in real time during live Cozmo interactions using human feedback, enabling contextually appropriate emotion display across six categories.

ROBOTIC PLATFORM

A small toy-like robot that can move,
make sounds and show face expressions

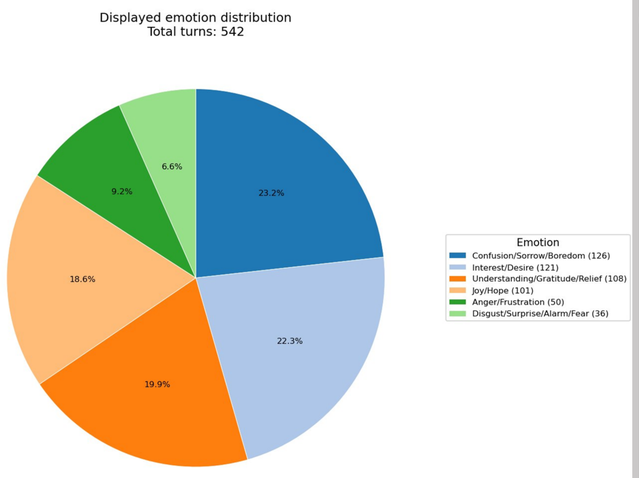
Cozmo



DATA

Synthetic Dataset (Experiment 1): 10,000 labeled dialogue samples generated using NVIDIA NeMo DataDesigner, simulating human questions paired with child-like Cozmo responses.

Real Interaction Dataset (Experiment 2): Collected during live human-Cozmo sessions across 18 sessions (~9 hours total), yielding 542 interaction turns with binary yes/no feedback, used simultaneously for real-time GRPO training.

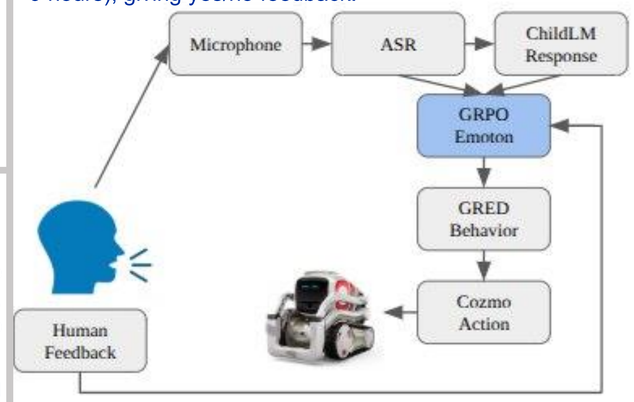


Emotion Categories: Both datasets share the same six emotion labels: *Anger/Frustration, Interest/Desire, Confusion/Sorrow/Boredom, Joy/Hope, Understanding/Gratitude/Relief, Disgust/Surprise/Alarm/Fear*

EXPERIMENTS

Experiment 1 (SFT + GRPO on Synthetic Data): DeBERTa-v3-base was fine-tuned on synthetic samples via SFT and further trained with GRPO using a composite reward, though live deployment on the Cozmo robot revealed frequent emotion mismatches with the interaction context.

Experiment 2 (Real-Time GRPO with Human Feedback): The SFT model was further trained with GRPO during live sessions. One tutor interacted across 18 sessions (542 turns, ~9 hours), giving yes/no feedback.



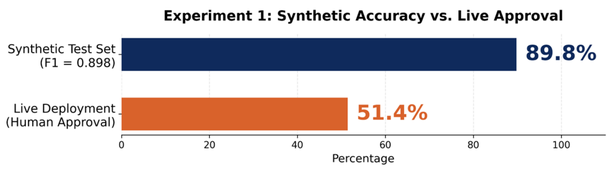
Exp. 3 (Human Evaluation): 10 new participants compared the Exp. 1 (baseline) and Exp. 2 (GRPO) models, blind and counterbalanced, then rated each on Godspeed and emotion-specific Likert items.

RESULTS

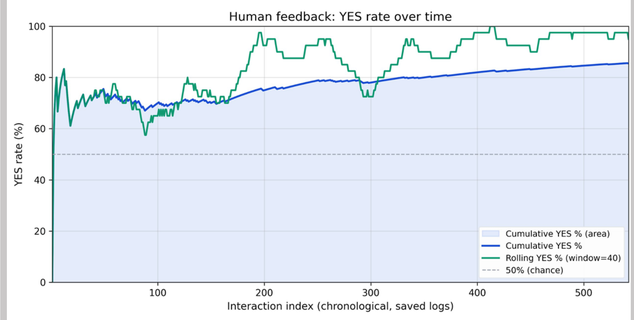
Experiment 1 (SFT + GRPO on Synthetic Data): The model achieved an accuracy of 89.79% and a macro-averaged F1 of 0.898 on the synthetic test set. Human evaluation revealed frequent emotion mismatches in the real interaction setting, motivating Experiment 2.

Per-class metrics

	Precision	Recall	F1	N
Anger	0.872	0.937	0.903	299
Interest	0.844	0.815	0.829	298
Confusion / sorrow	0.986	0.967	0.976	299
Joy	0.825	0.836	0.831	299
Gratitude / relief	0.911	0.923	0.917	299
Disgust / surprise / fear	0.954	0.909	0.931	298



Experiment 2 (Real-Time GRPO): Approval trended upward as the model received more live feedback, rising from 66.7% in the first session to 94.9% in the final session (85.4% overall), confirming that real-time GRPO successfully adapted the emotion policy to the target environment.

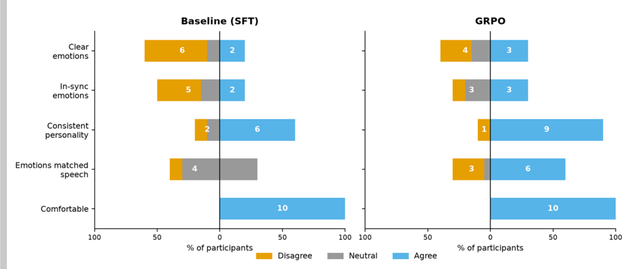


Experiment 3 (Human Evaluation): GRPO improved perceived emotion–speech alignment (6/10 agreed vs. 0/10 for baseline) and personality consistency (4.50 vs. 3.70). Comfort ratings were identical (10/10), confirming the gain was specific to emotional quality.

Godspeed-Style Ratings for Baseline (Exp. 1) vs. GRPO (Exp. 2) Models
Mean (SD), 1–5 Scale, N = 10 Participants

Item	Baseline	GRPO	% impr.
Confusing / Clear emotions	2.20 (1.23)	2.70 (1.42)	+22.7%
Out-of-sync / In-sync emotions	2.56 (1.33)	3.00 (1.15)	+17.2%
Inconsistent / Consistent personality	3.70 (1.16)	4.50 (0.97)	+21.6%

Participant Responses by Item: Baseline (Exp. 1) vs. GRPO (Exp. 2), N = 10



Emotion-Specific Likert Ratings for Baseline (Exp. 1) vs. GRPO (Exp. 2) Models
Agree/Neutral/Disagree Counts, N = 10 Participants

	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

CONCLUSIONS

- Online GRPO adapted the emotion policy in real time, raising approval from 51.4% to 85.4% (94.9% final session)
- Training remained stable throughout (bounded KL divergence, no drift)
- The adapted policy generalized to 10 new users, with clear gains in emotion-speech alignment (6/10 vs 0/10) and personality consistency (4.50 vs 3.70)

LIMITATIONS AND FUTURE WORK

Limitations: Real-time training relied on a single tutor, so the learned policy reflects one person's interpretation of appropriate emotion. Binary yes/no feedback is a coarse signal and a lab proxy — unsuited for real-world deployment. Ambiguous GRED behaviors and ChildLM speech sometimes confounded participants' ratings of emotional appropriateness.

Future Work: Future work will explore more naturalistic feedback signals, multi-user tutoring to reduce single-tutor bias, and tighter coupling between the emotion policy and GRED behavior generation for more context-aware, readable displays.

REFERENCES

[1] Retico: An incremental framework for spoken dialogue systems (Michael, SIGDIAL 2020)

[2] Recognizing and Generating Novel Emotional Behaviors on Two Robotic Platforms (Baral, Grenz, & Kennington, 2025)

[3] DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models (Shao et al., arXiv 2024)

[4] NeMo Data Designer Concepts and API Documentation (NVIDIA, 2025)

[5] Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots (Bartneck et al., 2009)

Anna Manaseryan: annamanaseryan@u.boisestate.edu