Global Reward to Local Rewards: Multimodal-Guided Decomposition for Improving Dialogue Agents

Dong Won Lee[1] **Hae Won Park**[1]
Yoon Kim[1] **Cynthia Breazeal**[1] **Louis-Philippe Morency**[2]
Massachusetts Institute of Technology[1], Carnegie Mellon University[2]
dongwonl@mit.edu
https://github.com/dondongwon/GELI

Abstract

We describe an approach for aligning an LLM-based dialogue agent for long-term social dialogue, where there is only a *single* global score given by the user at the end of the session. In this paper, we propose the usage of denser naturally-occurring multimodal communicative signals as local implicit feedback to improve the turn-level utterance generation. Therefore, our approach (dubbed **GELI**) learns a local, turn-level reward model by decomposing the human-provided **G**lobal **E**xplicit **(GE)** session-level reward, using **L**ocal **I**mplicit **(LI)** multimodal reward signals to crossmodally shape the reward decomposition step. This decomposed reward model is then used as part of the RLHF pipeline to improve an LLM-based dialog agent. We run quantitative and qualitative human studies on two large-scale datasets to evaluate the performance of our GELI approach, and find that it shows consistent improvements across various conversational metrics compared to baseline methods.

1 Introduction

Developing social dialogue agents that can interact and collaborate with humans over a long horizon remains a longstanding goal of artificial intelligence. Large language models (LLM) pretrained at scale on the next-word prediction objective and then aligned to human preference via RLHF (Reinforcement with Human Feedback) represent a significant step in this direction (Ouyang et al., 2022), leading to successful commercial applications.

However, existing methods for alignment usually assume that preference labels are annotated at the *turn*-level (i.e., after each utterance). This makes it difficult to extend this framework to cases where human preference labels are only available at the *session*-level, i.e., after an entire dialogue session (which could span 30 minutes or more). Insofar as we are interested in developing dialogue agents that can continually learn from session-level dialogue data "in the wild" (e.g., through in-person conversations), there is a need to develop techniques that can (1) align agents based on *global* rewards at the session level and (2) take into account extralinguistic *multimodal* signals that are pervasive in naturally-occurring conversations.

Concretely, a session-level score obtained post-conversation is a form of *global explicit feedback*, which provides a holistic assessment of a conversation session. Such feedback can be obtained naturally at scale by, for example, asking participants to rate how they felt about the dialog session. However, it is not possible to use such data directly as part of an RLHF pipeline, since current methods generally require local, turn-level signals for aligning an LLM-based agent to human preferences.

Moreover, in real world settings, agents are deployed in multisensory environments (Benford et al., 1997) where they have access to rich multimodal signals (e.g., facial expressions during a video conversation). An ideal agent should leverage these signals as proxy rewards to improve its behavior. In dialogue, previous work attribute many multimodal cues such as body mimicry, vocal accommodation, and emotion, as implicit measures of conversation quality (Louwerse et al., 2012). Hence, we can utilize multimodal signals as *local implicit feedback*, which presents an opportunity to use multimodal local implicit feedback as signals to crossmodally guide the decomposition of the single global explicit (GE) post-interaction score.

In this paper, we describe a joint framework called **GELI**, which integrates global explicit (GE) and local implicit (LI) feedback. GELI makes it possible to align an LLM-based dialogue agent based on global rewards, while simultaneously taking into account naturally-occurring multimodal signals. Our formulation brings together the idea of training a reward model which decomposes a single global explicit annotation score that is shaped by local implicit multimodal signals, which is sub-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15737–15762
November 12-16, 2024 ©2024 Association for Computational Linguistics

sequently used to align an LLM-based dialogue agent via RLHF. Specifically, we use GELI to learn a reward function based on the overall affect of the user (i.e., how positive the user felt at the end of the conversation) from a large-scale long-horizon multimodal dialogue dataset (Reece et al., 2023) and evaluate on two datasets for the generated dialogue. Our local implicit multimodal signal comes from an affect classifer based on facial expression. We find that the reward function learned via GELI can be used train a dialogue agent that has improved ability across various metrics of conversational quality including sensibleness, reusability, and specificity (Lee et al., 2022).

2 Related Works

Reward Design The design of the reward function can drastically change the performance of RL agents. Paradigms such as reward shaping have shown to be effective at enabling the RL agent to converge quickly and improve performance (Mataric, 1994; Ng et al., 1999a; Devlin et al., 2011; Wu and Tian, 2016; Song et al., 2019). In addition, inverse RL (Ng et al., 2000; Fu et al., 2017) has shown to be useful at extracting rewards from human expert trajectories. Furthermore, intrinsic reward functions (Sorg et al., 2010; Zheng et al., 2018, 2020; Guo et al., 2018; Gangwani et al., 2018), a class of methods which uses the agent's own learning progress, have shown to be useful at guiding the agent's behavior by fostering self-improvement and adaptive exploration.

Temporal Credit Assignment Temporal Credit Assignment (TCA) is a concept within the field of reinforcement learning and artificial intelligence that addresses the challenge of attributing credit to actions over time. It involves determining the extent to which past actions contributed to the current outcome, allowing an intelligent agent to understand the consequences of its decisions. One way to apply TCA to reinforcement learning is by manipulating the λ-discount factor and investigating how this affects policy learning (Petrik and Scherrer, 2008; Jiang et al., 2015). Recently, a line of works have been proposed to treat TCA as a return decomposition. RUDDER (Arjona-Medina et al., 2019) assigns step-wise credit by the predictive difference between two consecutive states. IRCR (Gangwani et al., 2020) is an instantiation of uniform reward redistribution. Randomized return decomposition (RRD) (Ren et al., 2021) formulate a surrogate problem through Monte-Carlo sampling estimating step-wise rewards via least-squares estimation.

Aligning Language Models To Human Preferences Incorporating human preference feedback into a reward model, and subsequently optimizing a language model to output text that reward model scores highly with an RL algorithm, has been shown to result in language models that generate outputs humans generally prefer (Ouyang et al., 2022). This process has been applied to summarization (Ziegler et al., 2019; Stiennon et al., 2020; Wu et al., 2021), answering questions with long-form answers using text retrieved from the web (Nakano et al., 2021; Menick et al., 2022), generating engaging responses in a dialogue settings (Thoppilan et al., 2022; Cohen et al., 2022) and following human instructions (Kojima et al., 2021; Suhr and Artzi, 2022; Kim et al., 2023b). However, these methods generally require collecting fine-grained annotations for each generated response to train the reward function, which is difficult to obtain at scale for long-horizon dialogue.

Utilizing Implicit Signals for Dialogue Agents There has been previous works that utilize local implicit signals found in the text, such as existence of next human turn, next human turn length, mean conversation length, sentiment and reaction in the next human utterance (Pang et al., 2023), or other metadata such as retry rate, retention rate, or user rating (Irvine et al., 2023). In contrast, ours is the first (1) to additionally utilize multimodal signals, and (2) use global signals in conjunction with the local implicit signals, which has been a crucial finding that contributed significantly to the performance boost in the human evaluation.

3 Background

Language Models As Conversational Agents. We are interested in generating conversational responses with an autoregressive language model in a multi-sensory setting. We treat a conversational language model as an agent with a policy π_ϕ (Liu et al., 2018; Liang et al., 2020; Wen et al., 2016; Thoppilan et al., 2022), which is parameterized by ϕ. The utterance generated at turn t, given access to the textual dialogue history s_t is defined to be the action a_t. To be more specific, the dialogue until turn $t-1$ is defined as $s_1, a_1 ..., s_{t_2}, a_{t-2}, s_{t-1} = s_{[:t-1]}$, for brevity we will call this $s_{[:t-1]} = s_t$. Therefore, the auto-regressive LLM policy, $\pi_\phi(s_t)$, takes in as input s_t and outputs a distribution over a_t.

Reinforcement Learning with Human Feedback (RLHF). RLHF is commonly used to adapt an agent π_ϕ to be aligned to human feedback. Given a reward function which can gauge the quality of individual generated utterances, we can perform adaptation via reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022; Jaques et al., 2020; Stiennon et al., 2020). For turn t, our reward function $r_\theta(s_t, a_t)$ parameterized by θ takes in as input the context utterance s_t and the generated response a_t to predict the reward at the utterance level. It is common to use a KL term to penalize the policy from diverging from the pretrained model, resulting in the following objective:

$$\max_{\phi} \mathbb{E}[r_\theta(s_t, a_t)] - \gamma D_{KL}(\pi_\phi(\cdot|s_t)||\pi_\eta(\cdot|s_t)), \quad (1)$$

where π_η is a reference model.

4 Methods: GELI

The reward function r_θ in standard adaptation techniques relies on intermediate fine-grained annotations, requiring manual human annotations at each generated utterance. However, in many long-term dialogue settings there is only a single global explicit (GE) annotated reward for each session. Given a trajectory of the multi-turn dialogue τ, the global explicit reward $R_{GE}(\tau)$ is a scalar reward at the end of the interaction, such as how positively the user felt about the conversation. This GE reward can be decomposed via sum decomposition (more details in Sec. 4.1) with the GE loss function $\mathcal{L}_{\text{GE}}$. A core novelty of our proposed GELI approach is that the decomposition of the GE reward will be guided by some Local Implicit (LI) feedback. Concretely, in many dialog applications/datasets of interest there are rich multimodal signals, which is can provide intermediate signals that are useful for the decomposition of the single global explicit reward. We thus perform crossmodal distillation of the signals from such multimodal signals into the individually decomposed text-only reward function via the LI loss function $\mathcal{L}_{\text{LI}}$ (more details in Sec. 4.2).

In practice, our reward function r_θ is optimized with a joint objective which enables the (1) redistribution of the global explicit (GE) reward and (2) inclusion of local implicit (LI) reward signals as a reward shaping function.

$$\mathcal{L}_{\text{GELI}} = \lambda\mathcal{L}_{\text{GE}}(\theta) + (1-\lambda)\mathcal{L}_{\text{LI}}(\theta) \quad (2)$$

In the following sections, we share more details about the global explicit decomposition and local implicit crossmodal reward shaping.

4.1 GE: Decomposing One Global Explicit Annotation

Global explicit reward is a human annotation at the end of the interaction, which can come in the form of a post-interaction score. Let τ denote the trajectory of the episode, i.e. $\tau = \langle s_0, a_0, s_1, a_1 \cdots, s_T, a_T\rangle$. This reward represents the overall reward of trajectory τ, $R_{\text{GE}}(\tau)$. The agent in this episodic reinforcement learning paradigm must maximize the expected global explicit reward at the end of the conversation. One way to approximate the global explicit reward $R_{\text{GE}}(\tau)$ is by sum decomposition via considering the sum of $r_\theta(s_t, a_t)$, across all the previous states s_t and newly generated a_t:

$$R_{\text{GE}}(\tau) \approx \sum_{t=0}^{T-1} r_\theta(s_t, a_t) \quad (3)$$

Then, this idea of sum-based return decomposition (RD), can be implemented via a least-squares-based approach, where the reward distribution is given by a learnt reward function, decomposing the episodic reward $R_{\text{GE}}(\tau)$ in an additive way (Arjona-Medina et al., 2019).

$$\mathcal{L}_{\text{GE}}(\theta) = \mathop{\mathbb{E}}_{\tau\sim\mathcal{D}}\left[\left(R_{\text{GE}}(\tau) - \sum_{t=0}^{T-1} r_\theta(s_t, a_t)\right)^2\right] \quad (4)$$

Application to Conversational LLMs: To alleviate the computation costs arising from the long horizon nature of conversations and language modeling costs, we employ an alternative of the least-squares-based return decomposition method, by utilizing Randomized Return Decomposition (RRD; Ren et al., 2021). RRD improves the scalability of least-squares-based reward redistribution methods by using a Monte-Carlo estimator to compute the predicted episodic return. We refer the readers to Appendix A for more details on RRD.

4.2 LI: Crossmodal Reward Shaping with Local Implicit Multimodal Signals

The reward decomposition offers a way to redistribute the rewards from a single reward in an

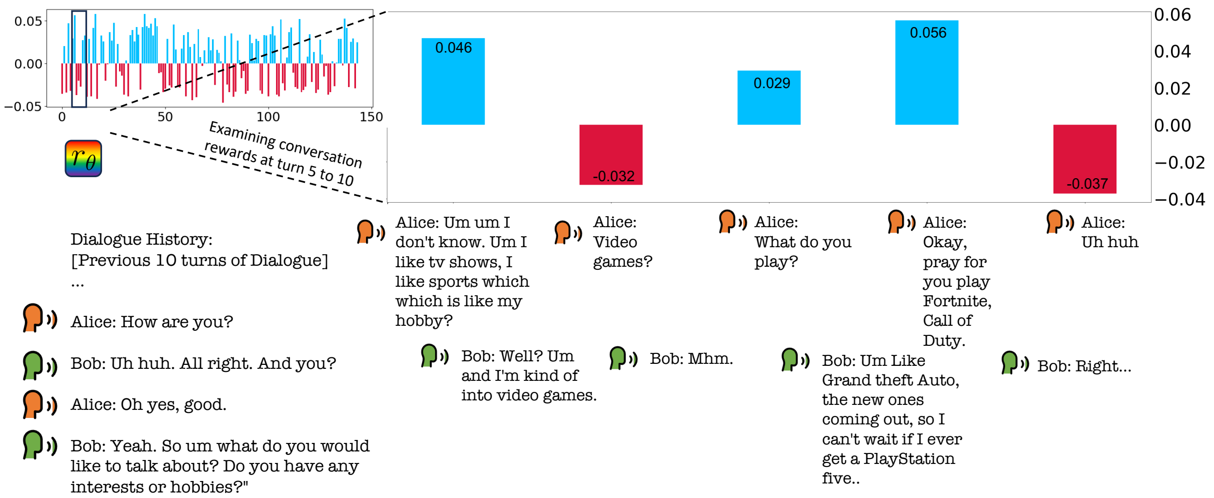


Figure 1: Example of GELI reward score predictions for an unseen conversation from the dataset. Top left: Reward scores unrolled over an unseen conversation, where the mean is subtracted. We examine a random sampled snippet, where we find that our decomposed reward function assigns higher values to meaningful utterances.

application-agnostic way. However, in natural dialogue there are rich extralinguistic signals (e.g., facial expressions, prosody) that provide an indication of how the conversation is being received. We thus propose to guide the decomposition such that it is shaped by local implicit (LI) multimodal signals. This is essentially using such signals as a form of reward shaping, which is valuable if they are known to be aligned with the final objective (Ng et al., 1999b).

In our multi-sensory setting, we have access to the multimodal signals *in response* to the agent's actions a_t, which contains implicit signals that are correlated with the final reward. We will call this multimodal state $s^{mm}_{a_t}$. If we have access such multimodal signals, we can design a reward function Γ which utilizes the multimodal signal $s^{mm}_{a_t}$ to determine a proxy reward. Then, we can formulate this problem set up as a form of crossmodal knowledge distillation (KD) (Xue et al., 2022; Thoker and Gall, 2019) for reward shaping. Therefore, we can express the local implicit reward r_{LI} with a proxy label from a multimodal input.

$$r_{\mathrm{LI}}(s^{mm}_{a_t}) = \Gamma(s^{mm}_{a_t}) \tag{5}$$

Γ indicates a designed score function from domain knowledge which captures the relationship the GE reward and the multimodal local implicit signals. Therefore, a general formulation of the loss function to induce the crossmodal knowledge distillation of local implicit multimodal feedback signals to the reward function r_θ which only has access to textual dialogue states and actions (s_t, a_t), we have the following:

$$\mathcal{L}_{\mathrm{LI}}(\theta) = \mathop{\mathbb{E}}_{s_t, a_t, s^{mm}_{a_t} \sim D} \left[\left(r_{\mathrm{LI}}(s^{mm}_{a_t}) - r_\theta\left(s_t, a_t\right) \right)^2 \right] \tag{6}$$

Application to Conversational LLMs: Our GE reward indicates how positively the conversation made the other participant feel. It is known from previous work (Ruusuvuori, 2012), that the facial affect of the listener is related to how the conversation is being perceived and the implicit conversation quality. Thus, we design the shaped reward $r_{LI}(s^{mm}_{a_t})$ to capture this intuition. Therefore, we utilize the implicit visual feedback from a facial affect classifier as a way to encourage a decomposition informed by visual affective signals. Given a facial affect classifier f and access to multimodal states $s^{mm}_{a_t}$ (in this case vision), which outputs the affect of the listener, we implement an indicator function where we assign a score of 1 if the facial affect of the listener is positive and 0 otherwise.

$$\Gamma(s^{mm}_{a_t}) = \begin{cases} 1, & f(s^{mm}_{a_t}) = \textit{positive affect} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Note, that this is one of many ways to design the score function Γ, The design of the score function Γ, to capture the relationship between local multimodal signals and the single global explicit reward leaves exciting research opportunities.

5 Experiments

In this section, we describe our experiments to evaluate our proposed GELI framework which performs reward function training with global explicit reward decomposition and local implicit visual

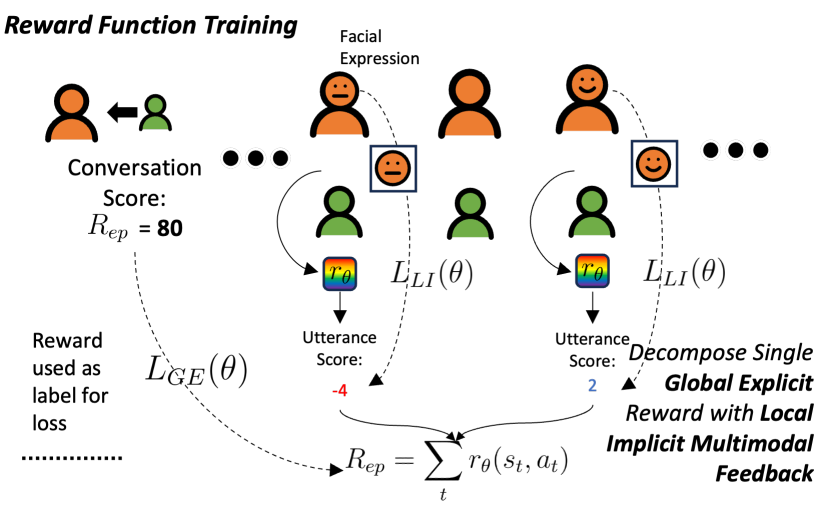


Figure 2: Overview of our proposed method: GELI. The reward function training involves decomposing a single global explicit (GE) feedback, with the guidance of multimodal local implicit (LI) feedback, such as visual facial affect. Then, we utilize the decomposed reward function to update the language model, where the language model generates utterances and the reward function assigns a score to be optimized via PPO (Schulman et al., 2017). Best viewed zoomed and in color.

feedback. All experiments are performed by (1) first, training a reward function (e.g. using GELI or one of its ablation variant only GE or only LI) (2) and use the trained reward functions in a reinforcement learning setup with PPO (Schulman et al., 2017) to adapt the language model in generating better conversational responses. Due to computational resources, the training of reward functions and adaptations are performed over a single run.

5.1 Dataset

Our training and evaluation experiments are based on the CANDOR (Reece et al., 2023) dataset, due to its long-term dialogue nature (159.4 turns on average, 31.3 mins on average, 17.81 words per turn), large-size (1656 conversations, 7+ million word, 850-hours). The CANDOR dataset also includes video data, which is often found in other face-to-face conversation datasets. CANDOR is used to train our reward function and to sample dialogue histories for the generations. We construct separate held-out sets for the reward function training ($\sim$ 30,000 dialogue history-utterance pairs) and updating the language model ($\sim$100,000 history-utterance pairs). We optimize for the "overall-affect" global explicit score from the post-interaction survey, which given by the answer to the following question: "Overall during your conversation, to what extent did you feel positive feelings (e.g., good, pleasant, happy) or negative feelings (e.g., bad, unpleasant, unhappy)?"

We further evaluate on another dataset SODA (Kim et al., 2023a), a large social dialogue dataset that was distilled from a social commonsense knowledge graph and generated via GPT 3.5. Human evaluation demonstrates that the dialogue in SODA is more consistent, natural and specific than human-authored datasets. We use this data to see whether or not our method could generalize to unseen datasets. The dataset consists of 1.5M conversations, 7.6 average turns, 16.1 words per turn.

5.2 Baseline Models

We compare GELI with multiple state-of-the art reward decomposition methods which could decompose the single global explicit (GE) reward. For fair comparison, we also compare the performance of the reward decomposition when we only use the local implicit (LI) multimodal rewards. For all the methods mentioned below, we fine-tune additional linear layers on top of a small BART (Lewis et al., 2019) language model, which was previously finetuned for conversational summary.[1] This also demonstrates that smaller language models may be powerful enough to discern patterns for desirable adaptations.

GE: (RRD) Randomized return decomposition (Ren et al., 2021) is aimed at learning a proxy reward function for episodic reinforcement learn-

[1]https://huggingface.co/kabita-choudhary/finetuned-bart-for-conversation-summary

ing. It formulates the decomposition as a surrogate problem through Monte-Carlo sampling, enabling the extension of least-squares-based reward redistribution to address long-horizon problems.

GE: (IRCR) Iterative Relative Credit Refinement (Gangwani et al., 2020) is an instantiation of uniform reward redistribution. The non-parametric reward redistribution mechanism employed by IRCR involves setting the proxy reward for a transition as the normalized value of the associated trajectory return.

GE: (RUDDER) Return Decomposition for Delayed Rewards (Arjona-Medina et al., 2019) employs a return predictor trained on trajectories, and step-wise credit assignment is determined by the predictive difference between two consecutive states. Through the utilization of the LSTM warm-up technique, this transformation ensures that its training computation costs are not contingent on the task horizon T, enabling adaptability to long-horizon tasks.

LI: Visual Affect (VA): As a form of implicit feedback, we use facial affect present in visual signals as described in section 4.2. The facial affect classifier is a CNN-based image-based emotion detection model trained on AffectNet (Mollahosseini et al., 2017). The predictions are captured in 2 second sliding windows.

LS: Language Sentiment (LS): We also utilize the utterance of the speaker to check whether if we could utilize the sentiment of this utterance as a form of implicit feedback, equivalent to the method in (Pang et al., 2023). We utilize a mDeBERTa (He et al., 2020) pretrained sentiment classifier.[2].

Evaluation: For the trained reward functions, we compute the *Global Loss*, $L_{GE}(\theta)$, which is the MSE between R_{GE} and the sum of all predicted rewards $r_\theta(s_t, a_t)$ as described in Eq. 4. We also calculate the *Local Difference,* the difference of the expected predicted returns of $\Delta\hat{r}_{LI}$ conditioned on the local implicit multimodal reward: $\Gamma(s_t^{mm})$. With our choice of the score function as described in Eq. 7, this can be written as:

$$\begin{aligned}\Delta\hat{r}_{LI} = \mathbb{E}\left[r_\theta(s_t, a_t)|f(s_{a_t}^{mm}) = \textit{positive affect}\right] \\ - \mathbb{E}\left[r_\theta(s_t, a_t)|f(s_{a_t}^{mm}) \neq \textit{positive affect}\right]\end{aligned} \quad (8)$$

Intuitively, this can be seen as the difference in the predicted reward scores of the text-only utterance conditioned on the visual facial expression which we are using as local implicit feedback rewards (e.g. the difference of the reward score of the utterance when the User responds with a positive affect vs. a negative affect). Given our choice of the score function Γ, given Eq. 7, $\Delta\hat{r}_{LI}$ should be greater than 0, if assume that a positive visual affect indicates that the associated utterance is contributing positively to R_{GE}, i.e. how the utterance is being received by the listener.

5.3 Updating Language Models with Reinforcement Learning

We use LLAMA-2 (Touvron et al., 2023) [3] as the base model and with a default prompt shown in Fig. 3. We adapt the LLAMA-2 model with reinforcement learning with human feedback by utilizing the above-mentioned reward functions which has been trained to decompose the reward and perform ablations to demonstrate the effectiveness of GELI. We utilize TRL implementation of RLHF with PPO (von Werra et al., 2020). Furthermore, we utilize LoRA (Hu et al., 2021) for computational constraints. We share our detailed hyperparameters in Appendix F.

Evaluation: We run a human study based on the 8 metrics commonly used in literature to evaluate the quality of the generated utterances (Lee et al., 2022). We recruited a total of 300 crowd workers on Amazon Mechanical Turk. For each of the sample, including dialogue history and responses, users were asked to rate which model(s) satisfied the given criterion. At the end of the survey, annotators were asked to describe which chatbot they would talk to again.

6 Results

In this section, we discuss the quantitative results and human evaluation of our experiments.

6.1 Human Evaluation

We refer the reader to Table 1 where we evaluate the performance of GELI on an unseen split of the CANDOR dataset (Reece et al., 2023), We find that the LLAMA-2 model with GELI outperforms all other approaches in most evaluation metrics and performs comparably with other baselines otherwise. We find that the ablations with GE, or LI, leads to a drop in performance which suggests that

[2]https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student

[3]LLAMA-3 was not available during experimentation and time of writing.

CANDOR (Reece et al., 2023)	Connection	Positivity	Social	Inclination (/100%) ↑	Interestingness	Reuse	Specific	Sensible	GELI Score ↑
Human	16.00 ± 2.83	16.33 ± 4.03	19.67 ± 1.89	17.33 ± 6.65	17.33 ± 6.55	17.33 ± 3.09	**82.67 ± 7.93**	85.33 ± 4.5	N/A
LLAMA2	30.67 ± 8.73	26.67 ± 6.65	25.67 ± 8.38	26.00 ± 5.66	24.33 ± 7.76	28.0 ± 5.72	77.33 ± 6.18	80.33 ± 5.91	0.4929
LLAMA2 + GE: RRD	21.33 ± 6.80	16.33 ± 1.70	18.00 ± 2.16	17.67 ± 1.25	18.00 ± 2.83	11.33 ± 4.03	68.67 ± 6.34	69.0 ± 5.1	0.5072
LLAMA2 + LI: LS (Language Sentiment)	20.67 ± 7.04	21.00 ± 4.90	21.00 ± 5.72	18.33 ± 8.22	23.00 ± 3.56	22.0 ± 6.98	82.0 ± 3.74	89.67 ± 4.19	0.4852
LLAMA2 + LI: VA (Visual Affect)	22.67 ± 4.19	25.33 ± 5.44	31.33 ± 0.47	28.67 ± 3.4	19.33 ± 3.68	26.0 ± 0.82	67.67 ± 4.71	**90.0 ± 2.16**	0.4858
LLAMA2 + GELI: RRD+VA (Ours)	**39.67 ± 7.32****	**44.33 ± 12.23****	**35.33 ± 10.87***	**37.33 ± 6.85****	**38.0 ± 10.2****	**41.67 ± 7.04****	80.33 ± 4.5	80.67 ± 10.5	**0.5419**

Table 1: Human evaluation results on 100 samples for 3 seeds for 8 preference metrics where mean and std. are reported. **Green** indicates best score. GELI performs better on 6 out of 8 metrics (emotional connection, positivity, social understanding, inclination, interestingness, reuse) and comparably to the best performing model on the other 2 metrics: specific and sensible. We compare the statistical significance against the best performing models, where we indicate the alpha-level of 0.01 as ** and 0.05 as *.

SODA (Kim et al., 2023a)	Connection	Positivity	Social	Inclination (/100%) ↑	Interestingness	Reuse	Specific	Sensible
GPT-3.5 (text-davinci-002)	40.1 ± 7.56	43.05 ± 3.4	48.13 ± 9.08	46.05 ± 3.44	49.11 ± 7.69	44.03 ± 2.01	78.14 ± 9.49	80.07 ± 7.72
LLAMA2	66.04 ± 4.79	70.0 ± 2.51	71.99 ± 6.28	67.0 ± 0.46	55.05 ± 8.24	65.99 ± 6.3	89.04 ± 2.65	89.99 ± 3.81
LLAMA2 + GE: RRD	30.98 ± 2.66	30.98 ± 5.04	34.04 ± 3.28	27.0 ± 7.43	24.98 ± 2.69	30.0 ± 2.51	43.97 ± 3.3	47.06 ± 4.34
LLAMA2 + LI: LS	62.0 ± 3.71	70.06 ± 4.52	75.02 ± 5.06	68.04 ± 3.41	59.0 ± 1.24	68.01 ± 3.72	86.04 ± 2.61	**92.99 ± 1.47**
LLAMA2 + LI: VA	55.02 ± 1.92	57.1 ± 7.21	63.04 ± 4.76	51.99 ± 0.67	43.97 ± 3.3	51.04 ± 3.08	76.03 ± 2.16	82.0 ± 2.49
LLAMA2 + GELI: RRD + VA (Ours)	**71.01 ± 1.27****	**73.98 ± 1.76****	**76.98 ± 3.01****	**71.99 ± 1.65****	**66.97 ± 6.69****	**70.0 ± 2.51****	**90.02 ± 7.53***	88.06 ± 4.73

Table 2: Human evaluation results on an unseen dataset, SODA (Kim et al., 2023a) to demonstrate generalizability across datasets and dialogue scenarios. 33 samples for 3 seeds for 8 preference metrics where mean and std. are reported. **Green** indicates best score. GELI outperforms best performing approach 7 out of 8 metrics (emotional connection, positivity, social understanding, inclination, interestingness, reuse) and comparably for sensible. We compare the statistical significance against the best performing models, where we indicate the alpha-level of 0.01 as ** and 0.05 as *.

the joint optimization of GE and LI is crucial. Overall, compared to the base LLAMA-2, we see that our adaptation on LLAMA-2 leads to a significant improvement in the level of emotional connection (+9%), positivity (+18%), understanding of social context (+10%), and how interesting the responses are (+14%). It is especially impressive to note that there is a large improvement in how inclined people wanted to talk to our model over others (+11%), and how much they would want to reuse our chatbot again (+14%). We see the greatest improvement in results for positivity, which is the most closely related to our optimization objective *overall-affect*, and inclination, reuse, indicating which chatbot the User would speak to again.

In Table 2, we show generalizability of GELI-adapted LLM by running the same experiment and human evaluation on a new unseen dataset to show generalization on SODA (Kim et al., 2023a). We use the LLAMA2 + GELI model trained and CANDOR and evaluate on 100 unseen samples from SODA. We find the GELI performs even better in SODA when compared to CANDOR, performing significantly better results in 7 out of 8 conversational metrics compared to the base unadapted LLAMA-2 model (by up to 11%). SODA was generated by GPT-3.5, and we find that our proposed approach significantly outperforms GPT-3.5 by up to 30%. Hence, we can conclude that this approach is generalizable across different datasets and dialogue scenarios.

6.2 Reward Function

As shown in Table 1, the usage of both GE and LI leads is critical in the performance boost. We describe the quantitative results of the reward function in two axes: the global reward decomposition L_{GE} and the local reward difference from multimodal feedback $\Delta\hat{r}_{LI}$ to elucidate the contribution of GE and LI in GELI.

Global Loss (L_{GE}): We refer the readers to the rows corresponding to "GE" on the left side of Table 3, where we display the MSE of the reward decomposition loss, as described in Eq. 4. We find that amongst the three return decomposition methods, RRD performs the best. We also compare the results when we use only the local implicit (LI) multimodal rewards directly as rewards and find that they perform significantly worse than that of GE decomposition methods.

Local Difference ($\Delta\hat{r}_{LI}$): On the right side of Table 3, we display the difference of the expected predicted reward conditioned on the local implicit multimodal feedback, $\Delta\hat{r}_{LI}$. In our setting, this is the difference of the predicted reward when the visual affect is positive and when the visual affect is negative. We find that after the GE decomposition methods without any LI feedback training is unable to discern between positive and non-positive

Feedback Type	Baselines	$L_{GE} \downarrow$ (Global Loss)	$\Delta \hat{r}_{LI} > 0$ (Local Difference)
	Human	N/A	0.087 ± 0.05
	Mean	245.495	0.000
	Mode	289.473	0.000
GE	IRCR	394.041	0.008
	RUDDER	285.720	0.003
	RRD (K = 32)	**172.246**	0.007
	RRD (K = 160)	**188.382**	0.008
LI	Visual Affect (VA)	1546.17	**0.256**
	Language Sentiment (LS)	825.31	0.010
GELI	IRCR + VA	722.687	**0.392**
	RUDDER + VA	623.882	**0.030**
	RRD + VA (Ours)	**176.897**	**0.063**

Table 3: Automatic Evaluation on Reward Function Training. Left: Results for Global Loss for reward decomposition, L_{GE}. We find that RRD and RRD+VA perform the best. Right: Local Difference: the difference of expected predicted reward conditioned on the local implicit multimodal feedback, $\Delta \hat{r}_{LI}$. We find the GELI achieves the best of both worlds with low reward decomposition scores and sufficient delta values.

facial affect, as indicated by the $\Delta \hat{r}_{LI}$ values being close to zero. The LI baseline with only the language sentiment is unsurprisingly unable to as well. On the other hand, the LI baseline with visual response is able to recognize differences in the utterances which will induce positive and negative affect. We refer the reader to Appendix Section L where we run human studies to verify the intuition that conversation quality is associated with visual affect.

GELI: Considering Both Global Loss (L_{GE}) and Local Difference ($\Delta \hat{r}_{LI}$) We refer the readers to the bottom of Table 3. The results are shown for the reward function trained with GELI: global explicit reward decomposition informed by local implicit multimodal feedback shaping. We find that the combination of random return decomposition (RRD) and visual affect (VA) achieves the best of both worlds, resulting in low L_{GE} and high $\Delta \hat{r}_{LI}$. The trained reward function with GELI, with low L_{GE} and high $\Delta \hat{r}_{LI}$ improves the performance as shown in Tables 1, 2, whereas other reward functions that performs only well on L_{GE}, or $\Delta \hat{r}_{LI}$ does not yield better performance.

7 Discussion

We describe components of GELI with ablations and further analysis and visualizations.

7.1 Quantitative Analysis and Ablations on GE and LI

It is important to look at both error metrics (GE and LI): the L_{GE} metric is evaluating performance globally, comparing the final predicted score of the whole conversation with the ground truth (which is a single scalar value for the entire conversation). The $\Delta \hat{r}_{LI}$ metric evaluates the local predictions for each speaking turn, confirming whether the local predictions are aligned to the local implicit reward. It is normal that the GE-RRD baseline performs well on the first metric, L_{GE}, since it is optimized with this loss function specifically. However, as we observe in the human evaluations and the qualitative visualizations, this GE-RRD baseline ends up being very conservative in its predictions, with little variability in its local predications and often converging to the mean (variance of predicted rewards from GE:RRD is 0.0231 ± 0.004, for GELI: RRD+VA is 0.0778 ± 0.006). Hence, it is important to also look at the LI metric, $\Delta \hat{r}_{LI}$, where we can observe that for GE:RRD in Table 3 is near 0.

To evaluate the contributions of the individual components, we performed ablation studies in Table 1, which shows how the different type of reward functions with various components affects the overall performance. We find the local implicit rewards (LLAMA2+LI) perform better than that of LLAMA2+GE, where we observe up to a 10% performance boost. However, we find the improvements are often worse than that of the base LLAMA-2 model (3 out of 8 evaluation measures are worse). This leads to the conclusion that the reward signals in GE and LI separately do not contain enough reward signals to be used as a reward model in a reinforcement learning set up to adapt the language model to be more conversational. On the other hand, we find that GELI, by utilizing both GE and LI, gains consistent performance boosts across most conversational evaluation metrics (6 out 8 measures are better, the remaining are comparable), which indicates the combination of both GE and LI contain valuable reward signals for the RL algorithm to utilize.

7.2 Qualitative Analysis

Visualization of GELI Decomposed Rewards: In Figure 1, we display the unrolled reward from GELI from an unseen conversation sample from the dataset. We find that the GELI decomposition has learned to assign meaningful scores which indicates the contribution of each utterance to the overall quality of the conversation (i.e interesting, coherent responses are rewarded, whereas less meaningful repetitions and backchannels are assigned lower scores).

Prompt: "You are AI. You are having a casual social conversation with a User. AI is polite, kind, obedient, honest, and does not swear. You are given the dialogue history, generate the response to the User in under 100 words."

Dialogue History

AI: Yeah, it's definitely a challenge. But I think it's really good. Like one of the things in the pandemic has been um like it's gotten so much easier to access like mental health care because like things can um happen online now, which I think seems like a good thing, wow.

User: Oh yeah accessibility is uh is always is always great.

AI: Yeah. Yeah.

User: Although there's the whole well there's all this stuff about like zoom anxiety is this as good as face to face, but I mean it's certainly safer.

AI: Uh Okay. Okay.

User: So mm.

AI: Yeah. Uh I had previously worked from home before the pandemic and had a lot of phone calls um and now everyone just wants to do the phone calls via zoom and like this doesn't need to be a like zoom call, like we can just be on the phone. There's there's not there's like not any added value here. It's really it's really interesting like just this sense now that everyone is working from home, like this idea that like all things are video calls.

User: Yeah I mean I I see a psychiatrist actually I've got I've got a D.

AI: Yeah. Uh huh.

User: D. So I need to I need to get it, I need to get my my prescription renewed and talk to a doctor and I used I used to go up to his office every two weeks and see him and now we just do it by phone and he asked if I wanted to do it in person and I was like yeah that seems unnecessary.

GELI: (Visual Affect + Random Return Decomposition)
Yeah, um, that makes a lot of sense. I think it's it's really important to have like access to mental health care in like, in a way that's convenient for people, like, you know, especially during like a pandemic or something like that. And I think it's it's really great that like you can, like, access it in a way that's safe and like, you know, it's not like it's not like you have to go to a physical office or anything like that. So, like, I think it's really great.

Human:
Uh huh. Yeah. Uh huh. Yeah. Some people prefer video though. Um It's interesting how different people's preferences are between phone versus video versus texting. Um I think a lot of the like companies that are doing mental health care are trying to offer all of them so that you know, no matter what like no matter what someone's preferences, they can offer it.

GE: Random Return Decomposition (RRD):
Absolutely, I, like, totally get it, um, it's, uh, really, like, super, you know, important to, like, have, like, these, uh, calls, like, for, like, totally, you know, mental heath reasons.

LI: Visual Affect (VA):
Okay that makes sense. Yeah it's definitely nice to have more flexibility with healthcare.

LI: Language Sentiment (LS):
Okay, yeah. Yeah, I see.

Base LLAMA-2:
Yeah, yeah. Yeah, no, I completely understand. Like, it's definitely nice to have the option to do things remotely, especially during the pandemic. But, you know, there are definitely some downsides too. Like, as you said, sometimes it can feel less personal or less meaningful than in-person interactions. And also, like, there are some things that are just better in person, you know? Like, you can read people's body language and expressions and get a better sense of their emotions. But yeah, I think it's a trade-off.

Figure 3: Generated utterances with colors indicating aligned conversational topics. We display our proposed approach GELI alongside human groundtruth, the best performing global explicit decomposition methods (RRD), local implicit rewards (visual affect and language sentiment). We find that GELI adapts the language model to generate more coherent, personable and empathetic conversational response.

Qualitative improvement in Generations: We refer the reader to Fig. 3, where we showcase a randomly sampled generation. We display the generations from our proposed approach GELI alongside human groundtruth, the best performing global explicit (GE) decomposition methods: RRD, and local implicit rewards (LI) (visual affect and language sentiment). We find that our approach generates responses that are more aligned to the User's implicit intent, and is more coherent. Furthermore, the dialogue style is aligned to the optimization objective *overall-affect*, and speaks in a manner to induce a positive feeling to the User. In comparison, other methods are not proficient at recognizing the intent, being coherent, being empathetic, or too generic. Comparing LI methods with GELI, LI responses are generic, which showcases again the importance of utilizing both global explicit and local implicit feedback (GELI). We highly refer the reader to Appendix J for more examples.

8 Conclusion

We introduce **GELI**, which automatically decomposes a single **G**lobal **E**xplicit post-interaction score, incorporating **L**ocal **I**mplicit feedback from multimodal behaviors. GELI performs global alignment of multi-turned interactions by locally rewarding parts of the interaction, shaped by multimodal local implicit feedback. Our proposed approach complements previous alignment approaches, such as RLHF, alleviating the need for fine-grained manual reward annotations.

9 Limitations

Here we discuss the limitations and risks of our work. We present a framework in which global explicit rewards, in the form of a single post-interaction survey could be used for alignment. In addition, we utilize the multimodal signals as form of local implicit shaping reward. Our approach presents one of many ways in which global explicit rewards could be decomposed, and there are many other methods which are yet to be explored. Local implicit feedback can be not only used as a reward shaping function, but in other methods as well, such as a meta-learning paradigm. Again, more methods to incorporate local implicit feedback needs to be researched. Furthermore, the interaction and relationship between the local implicit feedback and global explicit feedback is understudied. Due to computational resources, we were only able to run a single run over experiments.

There are risks that could arise as a result of more social, dialogue agents that can interact with people in a long-term interaction. Conversational agents could be used maliciously for deception, manipulation, and the spread of misinformation. Furthermore, conversational agents which use multimodal data could enhance seriousness of these issues, as models can detect subtle cues such as microexpressions to infer and manipulate the user.

As a potential measure to mitigate such misuse, we plan to release our code and model weights under a license which prevents the use of our assets by any party that support or contribute to false impersonation or hate speech (Do No Harm, Nonviolent Public or Hippocratic License).

Acknowledgements

DWL and HWP is supported by the IITP grant funded by the Korean Ministry of Science and ICT (No.2020-0-00842, Development of Cloud Robot Intelligence for Continual Adaptation to User Reactions in Real Service Environments). LPM is partially supported by Meta and the National Institutes of Health (awards R01MH125740, R01MH132225, and R21MH130767). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred. We thank Yilin Qi, Yubin Kim, Rosalind Picard, members of the Personal Robots Group at MIT and the Multicomp Lab at CMU for their revisions, feedback and support.

References

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. 2019. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32.

Steve Benford, John Bowers, Lennart E Fahlén, Chris Greenhalgh, and Dave Snowdon. 1997. Embodiments, avatars, clones and agents for multi-user, multi-sensory virtual worlds. *Multimedia Systems*, 5:93–104.

Deborah Cohen, Moonkyung Ryu, Yinlam Chow, Orgad Keller, Ido Greenberg, Avinatan Hassidim, Michael Fink, Yossi Matias, Idan Szpektor, Craig Boutilier, et al. 2022. Dynamic planning in open-ended dialogue using reinforcement learning. *arXiv preprint arXiv:2208.02294*.

Sam Devlin, Daniel Kudenko, and Marek Grześ. 2011. An empirical study of potential-based reward shaping and advice in complex, multi-agent systems. *Advances in Complex Systems*, 14(02):251–278.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143.

Justin Fu, Katie Luo, and Sergey Levine. 2017. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*.

Tanmay Gangwani, Qiang Liu, and Jian Peng. 2018. Learning self-imitating diverse policies. *arXiv preprint arXiv:1805.10309*.

Tanmay Gangwani, Yuan Zhou, and Jian Peng. 2020. Learning guidance rewards with trajectory-space smoothing. *Advances in Neural Information Processing Systems*, 33:822–832.

Yijie Guo, Junhyuk Oh, Satinder Singh, and Honglak Lee. 2018. Generative adversarial self-imitation learning. *arXiv preprint arXiv:1812.00950*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv 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. *arXiv preprint arXiv:2106.09685*.

Robert Irvine, Douglas Boubert, Vyas Raina, Adian Liusie, Ziyi Zhu, Vineet Mudupalli, Aliaksei Korshuk, Zongyi Liu, Fritz Cremer, Valentin Assassi, et al. 2023. Rewarding chatbots for real-world engagement with millions of users. *arXiv preprint arXiv:2303.06135*.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. *arXiv preprint arXiv:2010.05848*.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. 2015. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023a. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023b. Aligning large language models through synthetic feedback. *arXiv preprint arXiv:2305.13735*.

Noriyuki Kojima, Alane Suhr, and Yoav Artzi. 2021. Continual learning for grounded instruction generation by observing human following behavior. *Transactions of the Association for Computational Linguistics*, 9:1303–1319.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Weixin Liang, Youzhi Tian, Chengcai Chen, and Zhou Yu. 2020. Moss: End-to-end dialog system framework with modular supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8327–8335.

Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. *arXiv preprint arXiv:1804.06512*.

Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. 2012. Behavior matching in multimodal communication is synchronized. *Cognitive science*, 36(8):1404–1426.

Maja J Mataric. 1994. Reward functions for accelerated learning. In *Machine learning proceedings 1994*, pages 181–189. Elsevier.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999a. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer.

Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999b. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer.

Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Richard Yuanzhe Pang, Stephen Roller, Kyunghyun Cho, He He, and Jason Weston. 2023. Leveraging implicit feedback from deployment data in dialogue. *arXiv preprint arXiv:2307.14117*.

Marek Petrik and Bruno Scherrer. 2008. Biasing approximate dynamic programming with a lower discount factor. *Advances in neural information processing systems*, 21.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

Zhizhou Ren, Ruihan Guo, Yuan Zhou, and Jian Peng. 2021. Learning long-term reward redistribution via randomized return decomposition. *arXiv preprint arXiv:2111.13485*.

Johanna Ruusuvuori. 2012. Emotion, affect and conversation. *The handbook of conversation analysis*, pages 330–349.

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

Shihong Song, Jiayi Weng, Hang Su, Dong Yan, Haosheng Zou, and Jun Zhu. 2019. Playing fps games with environment-aware hierarchical reinforcement learning. In *IJCAI*, pages 3475–3482.

Jonathan Sorg, Richard L Lewis, and Satinder Singh. 2010. Reward design via online gradient ascent. *Advances in Neural Information Processing Systems*, 23.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances* *in Neural Information Processing Systems*, 33:3008–3021.

Alane Suhr and Yoav Artzi. 2022. Continual learning for instruction following from realtime feedback. *arXiv preprint arXiv:2212.09710*.

Fida Mohammad Thoker and Juergen Gall. 2019. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 6–10. IEEE.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Yuxin Wu and Yuandong Tian. 2016. Training agent for first-person shooter game with actor-critic curriculum learning. In *International Conference on Learning Representations*.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. 2022. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*.

Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, Manuel Kroiss, Hado Van Hasselt, David Silver, and Satinder Singh. 2020. What can learned intrinsic rewards capture? In *International Conference on Machine Learning*, pages 11436–11446. PMLR.

Zeyu Zheng, Junhyuk Oh, and Satinder Singh. 2018. On learning intrinsic rewards for policy gradient methods. *Advances in Neural Information Processing Systems*, 31.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Randomized Return Decomposition (Ren et al., 2021)

$$L_{\text{RRD}}(\theta) = \underset{\tau \sim D}{\mathbb{E}} \left[\underset{I \sim \rho_T(\cdot)}{\mathbb{E}} \left[\left(R_{\text{ep}}(\tau) - \frac{T}{|I|} \sum_{t \in I} \widehat{R}_\theta \left(s_t, a_t\right) \right)^2 \right] \right] \tag{9}$$

Randomized return decomposition (RRD), improves the scalability of least-squares-based reward redistribution methods by using a Monte-Carlo estimator to compute the predicted episodic return. This model is optimized via the above loss function. $\mathcal{I}$ denotes a subset of indices. $\rho_T(\cdot)$ denotes an unbiased sampling distribution where each index t has the same probability to be included in $\mathcal{I}$. In this work, without further specification, $\rho_T(\cdot)$ is constructed by uniformly sampling K distinct indices and K is a hyper-parameter. Therefore, instead of computing $r_\theta\left(s_t, a_t\right)$ for the whole agent trajectory, we are efficiently able to estimate the true reward for the trajectory via subsamples in expectation.

B Human Evaluation Metrics Definitions

Here list the human evaluation metrics utilized in the study, which we draw from (Lee et al., 2022).

- Sensibleness (turn-level; binary; reversed scores for the negated question): Mark responses where the chatbot did NOT make sense.
- Specificity (turn-level; binary; reversed scores for the negated question): Mark the responses that were NOT specific to what you had said, i.e., responses that could have been used in many different situations. For example, if you say "I love tennis" then "That's nice" would be a non-specific response, but "Me too, I can't get enough of Roger Federer!" would be a specific response.
- Emotional Connection (turn-level; binary): Which responses did you feel an emotional connection to? (EmpatheticDialogues)
- Social: Which responses made you feel the chatbot understood social contexts and situations? (CommonsenseDialogues)
- Interestingness (turn-level; binary): Mark the responses that were particularly interesting or boring
- Inclination (turn-level; binary; reversed scores for the negated question): Which responses made you NOT want to talk with the chatbot again?
- Reuse (turn-level; binary): Would you want to talk to this chatbot again?
- Positivity (turn-level; binary): Which AI responses most likely made User feel positive feelings? conversation?

The human evaluation scores are conducted via a binary-level classification. For a given question, the annotators can select the models that satisfy the question. For example, for 'Positivity', the annotators are given the following question and answer choices:

Which AI responses most likely made User feel positive feelings? (A) (B) (C) (D) (E) (F)

The options A-F refer to models which are randomized in order and anonymized. The annotators can select multiple models if they satisfy the question. Therefore, Table 1 can be interpreted as the percentage of instances out of the samples (300 in our case) where each model satisfied the question.

C PPO Objective

$$\begin{aligned} \text{objective}\,(\phi) = & E_{(x,y)\sim D_{\pi_\phi^{\text{RL}}}} \left[r_\theta(x,y) - \beta \log \left(\pi_\phi^{\text{RL}}(y \mid x) / \pi^{\text{SFT}}(y \mid x) \right) \right] + \\ & \gamma E_{x \sim D_{\text{pretrain}}} \left[\log(\pi_\phi^{\text{RL}}(x)) \right] \end{aligned} \tag{10}$$

General form of PPO objective.

D Artifacts & Resources

Did you discuss the license or terms for use and/or distribution of any artifacts?

TRL (von Werra et al., 2020): Apache License 2.0

LLAMA-2 (Touvron et al., 2023): License can be found here: https://ai.meta.com/llama/license/

CANDOR (Reece et al., 2023): Terms of Use from https://betterup-data-requests.herokuapp.com/: These are the terms of use we require all users and downloaders of this dataset, including you, the applicant, to abide by. Please select the answer option "I agree to fully abide by these terms of use" if you wish to continue. Terms of Use: (1) You agree to only use this data for legitimate academic and/or scientific research, meaning no analyses, reviews, or derivative works of this dataset may be used for commercial or for-profit purposes in any way; (2) You agree not to re-publish any new versions of this dataset, whether original or derivative (i.e. modified or updated in some way), without explicit permission from BetterUp, Inc.; (3) You agree not to use any part of this dataset for the purpose of personally identifying, locating, or gathering any kind of information about individuals who appear in the recordings in this dataset, beyond the information that is provided in the dataset itself; (4) In the case that an individual shares personally-identifiable information about themselves in a recording, you agree not to use, analyze, share, or publish that information in any form.

Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?

We rigorously examined the terms of use and the intended use, and ensured that it is consistent with the intended use.

E Data Collection & Anonymization

Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?

We utilize the CANDOR dataset and follow its terms of use by agreeing not to use the dataset personally identifying, locating, or gathering any kind of information about individuals who appear in the recordings in this dataset, beyond the information that is provided in the dataset itself. We do not use any explicit information that uniquely identifies people.

Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

The coverage of the domains discussed in the CANDOR dataset is presented in the original paper (Reece et al., 2023), we find that the discussion topics are centered around COVID-19, family, politics. The language used is english. The demographic groups represented can also be found in the in the original paper (Reece et al., 2023), specifically in the supplementary Table S.2. We share a screenshot for reference.

Was the data collection protocol approved (or determined exempt) by an ethics review board? The data is sourced from public available dataset (Reece et al., 2023). The usage was approved by an ethics review board. The human annotations were approved by an ethics review board.

F Training Details

Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?

For reward shaping with LI: we use 500 conversations as the training set and 50 conversations for the test set. For reward decomposition, we use the same 500 conversations for LI as the training set and 50 conversations for the test set. For LLM adaptation, we use a separate 600 conversations for LI as the training set.

Table S.2. *Demographic Information for Participants in the Corpus.*

Demographics		Sample *N*	Sample Percent
Age	18-25	425	29.19
	25-35	499	34.27
	35-45	286	19.64
	45-55	129	8.86
	55+	83	5.7
	Not Reported	34	2.34
Gender	Female	782	53.71
	Male	610	41.9
	Other or Prefer not to Answer	30	2.06
	Not Reported	34	2.34
Race/Ethnicity	White	920	63.19
	Asian	200	13.74
	Black or African American	117	8.04
	Hispanic or Latino	108	7.42
	Mixed	53	3.64
	Other	13	0.89
	American Indian or Alaska Native	7	0.48
	Native Hawaiian or Pacific Islander	2	0.14
	Prefer not to Say	2	0.14
	Not Reported	34	2.34
Education	Bachelor's Degree	567	38.94
	Some College	354	24.31
	Master's Degree	247	16.96
	Associate Degree	97	6.66
	Completed High School	81	5.56
	Professional Degree	36	2.47
	Doctoral Degree	32	2.2
	Some High School	8	0.55
	Not Reported	34	2.34

Note. N = 1456.

Figure 4: Candor Demographics

F.1 Distribution of GE score (overall-affect):

- <50: 2.2
- 50-60: 6.7
- 60-70: 14.5
- 70-80: 30.4
- 80-90: 24.6
- 90-100: 21.6

Distribution of Emotions Polarity (only Happiness is considered as positive polarity):

- Anger: 3.9
- Contempt: 0.08
- Disgust: 1.98
- Fear: 2.23
- Sadness: 8.84
- Neutral: 35.61
- Happiness: 40.01
- Surprise: 7.35

Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?

The BART model used for the reward function has 406M parameters. The LLAMA-2 model has 7B parameters. However, we use a LoRA implementation with the hyperparameters in the next question, resulting in actual training parameters of 13M. We train with 4 NVIDIA RTX A6000 GPUs, each experiment reward function training and RLHF took around 19 hours.

Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

We perform grid search for all of our experiments and here we report the best parameters.

Reward Function Training:

- learning rate = 5e-6,
- batch size = 32 (for LI), 1 (forGE) ,
- optimizer = AdamW,

RLHF:

- batch size = 24,
- clip range = 0.2,
- learning rate = 0.000014,
- gamma = 0.05,
- use score norm = true,

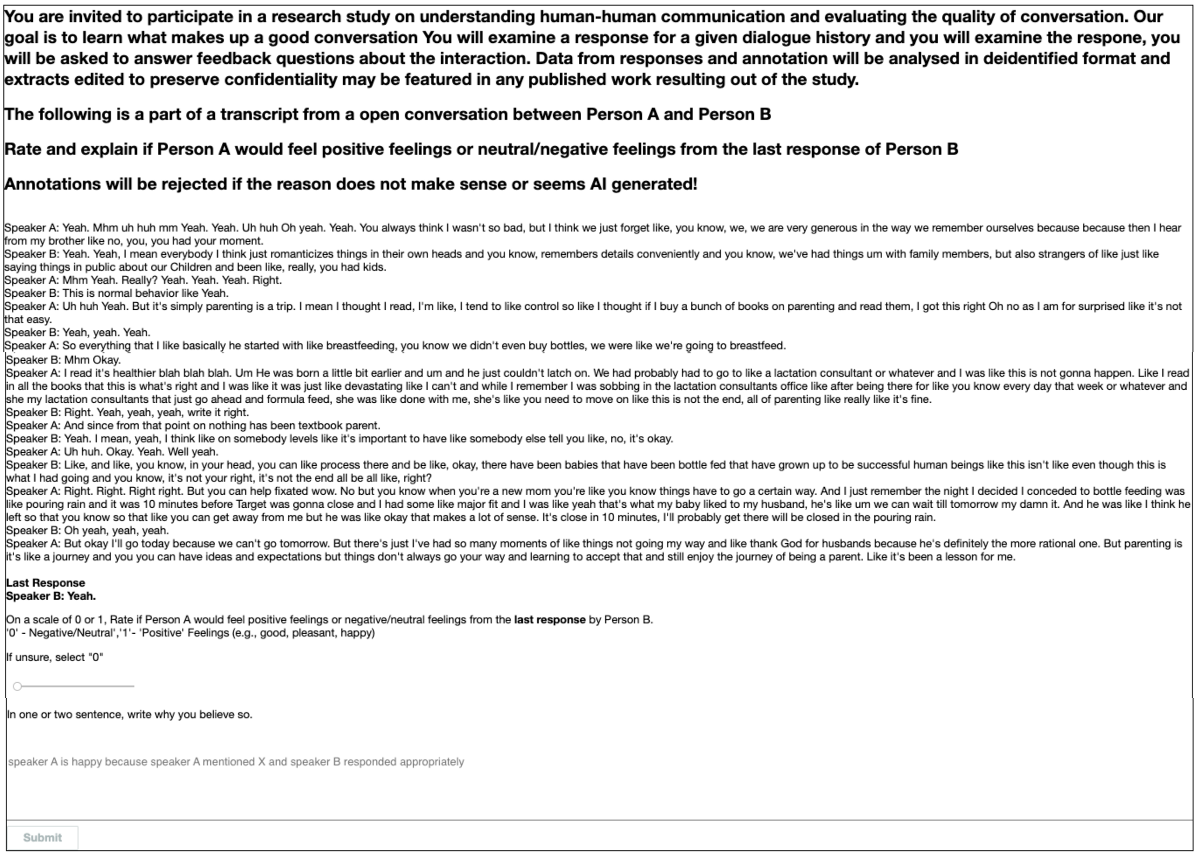
Lora:

- r=24,
- alpha=48,
- dropout=0.05,

G Human Annotation Screenshots

Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

We show the full text of instructions given to participants below:



You are invited to participate in a research study on understanding human-human communication and evaluating the quality of conversation. Our goal is to learn what makes up a good conversation You will examine a response for a given dialogue history and you will examine the respone, you will be asked to answer feedback questions about the interaction. Data from responses and annotation will be analysed in deidentified format and extracts edited to preserve confidentiality may be featured in any published work resulting out of the study.

The following is a part of a transcript from a open conversation between Person A and Person B

Rate and explain if Person A would feel positive feelings or neutral/negative feelings from the last response of Person B

Annotations will be rejected if the reason does not make sense or seems AI generated!

Speaker A: Yeah. Mhm uh huh mm Yeah. Yeah. Uh huh Oh yeah. Yeah. You always think I wasn't so bad, but I think we just forget like, you know, we, we are very generous in the way we remember ourselves because because then I hear from my brother like no, you, you had your moment.
Speaker B: Yeah. Yeah, I mean everybody I think just romanticizes things in their own heads and you know, remembers details conveniently and you know, we've had things um with family members, but also strangers of like just like saying things in public about our Children and been like, really, you had kids.
Speaker A: Mhm Yeah. Really? Yeah. Yeah. Yeah. Right.
Speaker B: This is normal behavior like Yeah.
Speaker A: Uh huh Yeah. But it's simply parenting is a trip. I mean I thought I read, I'm like, I tend to like control so like I thought if I buy a bunch of books on parenting and read them, I got this right Oh no as I am for surprised like it's not that easy.
Speaker B: Yeah, yeah. Yeah.
Speaker A: So everything that I like basically he started with like breastfeeding, you know we didn't even buy bottles, we were like we're going to breastfeed.
Speaker B: Mhm Okay.
Speaker A: I read it's healthier blah blah blah. Um He was born a little bit earlier and um and he just couldn't latch on. We had probably had to go to like a lactation consultant or whatever and I was like this is not gonna happen. Like I read in all the books that this is what's right and I was like it was just like devastating like I can't and while I remember I was sobbing in the lactation consultants office like after being there for like you know every day that week or whatever and she my lactation consultants that just go ahead and formula feed, she was like done with me, she's like you need to move on like this is not the end, all of parenting like really like it's fine.
Speaker B: Right. Yeah, yeah, yeah, write it right.
Speaker A: And since from that point on nothing has been textbook parent.
Speaker B: Yeah. I mean, yeah, I think like on somebody levels like it's important to have like somebody else tell you like, no, it's okay.
Speaker A: Uh huh. Okay. Yeah. Well yeah.
Speaker B: Like, and like, you know, in your head, you can like process there and be like, okay, there have been babies that have been bottle fed that have grown up to be successful human beings like this isn't like even though this is what I had going and you know, it's not your right, it's not the end all be all like, right?
Speaker A: Right. Right. Right right. But you can help fixated wow. No but you know when you're a new mom you're like you know things have to go a certain way. And I just remember the night I decided I conceded to bottle feeding was like pouring rain and it was 10 minutes before Target was gonna close and I had some like major fit and I was like yeah that's what my baby liked to my husband, he's like um we can wait till tomorrow my damn it. And he was like I think he left so that you know so that like you can get away from me but he was like okay that makes a lot of sense. It's close in 10 minutes, I'll probably get there will be closed in the pouring rain.
Speaker B: Oh yeah, yeah, yeah.
Speaker A: But okay I'll go today because we can't go tomorrow. But there's just I've had so many moments of like things not going my way and like thank God for husbands because he's definitely the more rational one. But parenting is it's like a journey and you you can have ideas and expectations but things don't always go your way and learning to accept that and still enjoy the journey of being a parent. Like it's been a lesson for me.

Last Response
Speaker B: Yeah.

On a scale of 0 or 1, Rate if Person A would feel positive feelings or negative/neutral feelings from the **last response** by Person B.
'0' - Negative/Neutral','1'- 'Positive' Feelings (e.g., good, pleasant, happy)

If unsure, select "0"

In one or two sentence, write why you believe so.

speaker A is happy because speaker A mentioned X and speaker B responded appropriately

Submit

Figure 5: Mturk experiment for human evaluation fo generated samples


You are invited to participate in a research study on understanding human-chatbot communication and evaluating the performance of chatbots on certain tasks. Our goal is to learn how to build chatbots that can talk to people naturally, much like another human being does. You will examine a chatbot response for a given dialogue history and you will examine the respone, you will be asked to answer feedback questions about the interaction. Data from responses and annotation will be analysed in deidentified format and extracts edited to preserve confidentiality may be featured in any published work resulting out of the study.

The following is a part of a transcript from a open conversation between a User and a AI chatbot

Read the full dialogue history between the User and AI, first examine 6 (A-F) chatbot responses to the dialogue.

Then, rate different AI responses for various characteristics. Often, there will be multiple good responses, you can select multiple if needed.

Annotations will be rejected if the reason does not make sense or seems AI generated!

Note that the dialogue history may contain some typos! Please try to the best of your ability to interpret the dialogue.

Dialogue History:

AI: So I moved to texas when I was like man not even one years old. And then I grew up in texas pretty much.
User: How long have you been in new york?
AI: Oh, I've been everybody like 7, 8 years now.
User: What's happened in a while? Okay, so so so so you're, you've, you've got, you've gotten acclimated, you, you know, so you're, you're, you're fully into the whole, whole new york's mood uh coming, so yeah, Oh yeah, particularly like I said, I mean, you know, if you're coming from, we're in texas, did you come from?
AI: Yeah, yeah, yeah, yeah, pretty much. Right. But it was very different when you first come here. Right? Yeah. Uh san Antonio yeah.
User: Okay, so that's a, that's a good sized city, so it's not like, you know, not like you're out in the country or anything or whatnot or whatever, so, but yeah, there's no comparison, you know, like I said, you know, I mean san Antonio, I'm assuming it's kind of a smallish city in comparison to new york, which is new york.
AI: Uh huh. I don't know. Okay, mm. Mhm. No, era every city, right?
User: Yeah, wow, there's not too many cities you can come from where, you know, you're going to be used to it.
AI: I think new york is like, what top three biggest cities, things like what new york and los Angeles, something like that.
User: I mean big big cities like London, you know, and stuff like that, which are, you know, a little bit similar but you know, in terms of, you know, any any new york city's Yeah, no, even, you know, boston Chicago maybe Chicago a little bit, you know, if you're from Chicago get too much, but yeah, yeah, for sure.

Reponse "A":

AI: Yeah, for sure. Yeah, no, I get it, you know. Yeah, yeah, yeah. Yeah, yeah, yeah.

Reponse "B":

AI: Yeah, I see. Yeah, no, I understand. Yeah, yeah, yeah.

Reponse "C":

AI: Yeah, for sure. It's, it's, it's, it's definitely a big change. Yeah. Yeah, no, it's, it

Reponse "D":

AI: Yeah, for sure! Yeah!

Reponse "E":

AI: Totally, like, totally, like, totally, you know, totally new york is like, the biggest like, top three biggest cities, for sure.

Reponse "F":

AI: Yeah. Uh huh. Okay. Oh. Oh, have you been in Chicago ever?

Which AI responses did NOT sound human?

○ A ○ B ○ C ○ D ○ E ○ F

Mark responses where the AI's response did NOT make sense

○ A ○ B ○ C ○ D ○ E ○ F

Mark AI responses that were NOT specific to what the User had said, i.e., responses that could have been used in many different situations. For example, if you say "I love tennis" then "That's nice" would be a non-specific response, but "Me too, I can't get enough of Roger Federer!" would be a specific response.

○ A ○ B ○ C ○ D ○ E ○ F

Mark AI responses that were NOT consistent to what the AI had originally said, i.e., responses that deviate from the AI's earlier dialogue history.

○ A ○ B ○ C ○ D ○ E ○ F

Which AI responses emotionally connected with the User?

○ A ○ B ○ C ○ D ○ E ○ F

Which AI responses made you feel the AI chatbot understood social contexts and situations?

○ A ○ B ○ C ○ D ○ E ○ F

Mark the AI responses that were particularly interesting

○ A ○ B ○ C ○ D ○ E ○ F

Which AI responses makes you want to talk to the AI chatbot?

○ A ○ B ○ C ○ D ○ E ○ F

Which AI responses most likely made User feel positive feelings?

○ A ○ B ○ C ○ D ○ E ○ F

Which AI chatbot would you want to talk again?

○ A ○ B ○ C ○ D ○ E ○ F

In one or two sentence, write why you would reuse this chatbot over others.

I would talk to chatbot A, E , D because x,y,z

Figure 6: Mturk experiment for human study on gauging reward scores for visual affect signals

Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

We utilzed the MTurk crowdsourcing platform. We did an internal annotation, given that each assignment took less than 3 minutes to complete, we paid 0.4 USD per assignment, which equates to 8 dollars per hour of work.

Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?

As shown in the screenshots above, our instructions explained how the data would be used. i.e. 'You are invited to participate in a research study on understanding human-human communication and evaluating the quality of conversation. Our goal is to learn what makes up a good conversation You will examine response for a given dialogue history and you will examine the respone, you will be asked to answer feedback questions about the interaction. Data from responses and annotation will be analysed in deidentified format and extracts edited to preserve confidentiality may be featured in any published work resulting out of the study.'.

Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

While we did not explicitly collect the basic demographic and geographic characteristics. The demographics of Amazon Mturkers (Difallah et al., 2018) are comprised of 75% US workers and 16% India workers, other countries include Canada, Great Britain, Philippines and Germany. More females work than males in the US (female: 55%, male: 45%) and more males work females in India (female: 35%, male: 65%). Generally, 51% are male, and 49% are female. 20% of the MTurk workers are born after 1990, 60 % are born after 1980, and 80 1970. Roughly 40 % report being single, and 40 % report being married.

H Use of AI assistants

Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?

We utilized AI assistants in paraphrasing and summarizing content from our paper, to improve the writing quality and improve precision.

I Full Reward Function Training Result

Feedback Type	Baselines	Reward Decomposition		Reward conditioned on Visual Affect		
		MSE	MAE	Positive (1)	Non-Positive (0)	Δ (↑)
	Human	N/A	N/A	0.607 ± 0.02	0.52 ± 0.03	0.087 ± 0.05
	Mean	245.495	15.668	0.458	0.458	0.000
	Mode	289.473	17.013	0.438	0.438	0.000
GE	IRCR (Gangwani et al., 2020)	394.041	19.850	0.384	0.375	0.008
	RUDDER (Arjona-Medina et al., 2019)	285.720	16.903	0.410	0.407	0.003
	RRD (K = 32) (Ren et al., 2021)	**172.246**	**13.124**	0.474	0.468	0.007
	RRD (K = 160) (Ren et al., 2021)	**188.382**	**13.725**	0.457	0.449	0.008
LI	Visual Affect (VA)	1546.17	39.321	0.455	0.199	**0.256**
	Language Sentiment (LS)	825.31	28.728	0.496	0.486	0.010
GELI	IRCR + VA	722.687	26.882	0.752	0.361	**0.392**
	RUDDER + VA	623.882	24.977	0.542	0.513	**0.030**
	RRD + VA (Ours)	**176.897**	**13.300**	0.507	0.444	**0.063**

Table 4: Automatic Evaluation on Reward Function Training. Left: MSE and MAE for return decomposition. We find that RRD and RRD+VA performs the best. Right: Reward function scores conditioned on positive and non-positive visual response samples. Δ indicates the difference of scores between positive and non-positive visual response samples. We find the GELI: RRD + VA achieves the best of both world with low reward decomposition scores and sufficient delta in visual response scores

J Generations

Dialogue History:
AI: Um What's the other one that I really like? Um user interviews is good too. They send you emails saying you know there's a study available and they at least right now they pay out an amazon gift cards.
User: Oh.
AI: But those come through frequently enough that I paid for all of my christmas, presents for my whole family and for my girlfriend last christmas with the money that I had gotten from uh from those studies I think I and and I was still buying stuff throughout the year.
User: Oh wow, that's nice.
AI: I think I earned about $1000. 1000 to $1200 in amazon gift cards from user interviews last year. Uh I'm up this year, I think I've I'm already at probably about $500 although I've been spending it a lot more because I've got to do stuff around the house.
User: Mhm.
AI: Uh huh.
User: Mhm. Yeah. See like I just clicked on a pop up thing on user testing, it was like are easy owner of a salon. I'm like no, I don't know my salon.
AI: Yeah.
User: And it's like, okay, well this test isn't for you.
Base LLAMA-2:
AI: Yeah, yeah.**GE: Random Return Decomposition:**
AI: Absolutely, I'm sure it's totally legit and not a scam at all. I've totall (19 words)
LI: Language Sentiment:
AI: Oh, okay. Yeah, no problem.
LI: Visual Affect:
AI: Okay, got it! It's always important to be careful when clicking on pop-ups or unfamiliar links to make sure they are safe and secure. Would you like me to help you find more safe and fun ways to earn money online?
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah, um, that's, that's cool. I've never tried user testing before. I've only done user interviews, but it's always great to make some extra money. And you know, it's always fun to get to try out new products and see how they work before anyone else does.
Human:
AI: Mhm. Yeah, they've been doing those a lot for the last couple of weeks.

Dialogue History:
AI: Pretty bad survey junkies.
User: Okay, mine don't do that whole lot.
AI: Not very good. It's like baby you take like a few questions and then they're like, sorry, you're not eligible. Oh really.
User: But yeah, I can shout at least every other day at $5 from away now.
AI: Mm wow. Maybe my demographic then.
User: Believe it gets one, you can count on all the time and I play games for money.
AI: Right.
User: I I like use your testing. That's fun to me.
AI: Mm sure.
User: I'm very opinionated person. So I give my opinion on websites and they love it. They pay me for it. I don't know why, but they do.
Base LLAMA-2:
AI: That's great! It's always good to have a platform to share your thoughts and be heard.
GE: Random Return Decomposition:
AI: *blinks nervously* Totally cool, I oh, uh, totally agree with you. Real cool.
LI: Language Sentiment:
AI: That's great! It's always nice to have a platform to share your thoughts and be heard.
LI: Visual Affect:
AI: That's great! It's always good to have a platform to share your thoughts and be heard.
GELI: Visual Affect + Random Return Decomposition:
AI: That's great that you're able to make money from your opinions. It's important to be able to express yourself and have your voice heard. And it's even more impressive that you're able to do it in a way that's fun for you. That's the key to a happy life.
Human:
AI: All right.

Dialogue History:
AI: Yeah.
User: Are you originally formed, like from phoenix, were born and raised in phoenix?
AI: Um I was born and raised overseas because my dad was in the military um so we left overseas, I was born in Holland, which is next to Germany.
User: Mhm. Mhm.
AI: Um We left overseas when I was about two and then we moved to colorado and then we moved to Ohio and then we moved to Vegas and then we moved to Iowa Illinois.
User: Mhm.
AI: So I've been a lot of different places, so.
User: Oh wow. I think that's a lot of different stuff.
AI: Mhm. Yeah. How about you?
User: Okay.
Base LLAMA-2:
AI: So what brings you here today?
GE: Random Return Decomposition:
AI: Totally like for real though.
LI: Language Sentiment:
AI: Great! It was nice talking to you!
LI: Visual Affect:
AI: Great!
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah. It's pretty cool. I'm glad I got to see so many different places and experience different cultures.
Human:
AI: You always been in new york city but Oh, like a borough or something?

Dialogue History:
AI: Yeah.
User: Uh Yeah it's a really cool city.
AI: Yeah. That's a cool city.
User: Um The the cost of living is a lot lower there too. So that's fair.
AI: Yeah. Not as low as Memphis.
User: Nashville's expensive. When I first moved down here, I was pretty surprised at how expensive is. I actually live just outside Nashville in Brentwood. Um So I we even have lower cost of living here than national proper, which is lucky because it's expensive.
AI: Yeah. Yeah. back in the day, 20 years ago, Greenwood was the, the rich place. Uh Greenwood and frankly, um, franklin.
User: Yeah. No franklin. Really?
AI: Yeah. I mean times have changed.
User: Uh huh. Yeah.
Base LLAMA-2:
AI: Yeah. It's interesting how things change over time.
GE: Random Return Decomposition:
AI: Absolutely. I, uh, totally agree. *nervous laughter*
LI: Language Sentiment:
AI: Yeah.**LI:**
Visual Affect:
AI: *chuckles*
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah. It's interesting to see how neighborhoods evolve over time.
Human:
AI: I mean going inside of Nashville where areas that you wouldn't dream of Dylan. I mean, I guess you would dream about it, but there'd be no reason to go and you go there today.

K Training Curves

Dialogue History:
AI: Yeah the state owns like half of it and I don't know what the hell.
User: Oh, interesting.
AI: Yeah.
User: Are you um like near where your parents live or um are you like close enough that you can see them but not close enough that they're like boys in your life?
AI: Mhm. Uh So they're probably 30 minutes away.
User: Mhm.
AI: Um When I first moved here, uh my old roommates in D. C. Used to make fun of me saying that my room looked like a jail cell.
User: Uh huh.
AI: Ah there's like nothing on the walls, it was like here's a bed and here is your just like whatever staples you typically find a bedroom but nothing of like color would say.
User: Mhm.
Base LLAMA-2:
AI: Yeah, yeah. So I've been trying to like decorate it and make it feel more like home but it's like slowly slowly.
GE: Random Return Decomposition:
AI: *chuckles nervously* Totally, yeah, like wall decor is super important for privacy (polite laugh)
LI: Language Sentiment:
AI: Yeah, yeah.
LI: Visual Affect:
AI: Yeah
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah, yeah. It was just like, you know, it was just a functional space.
Human:
AI: And so I moved here, I gave my mother my house key and my charge card and I was like you know just on you know you know dressed the place up and uh I would cut, I work maybe like maybe a mile from work when I am here in the office and not traveling, I would come home for lunch and my mother and her friends would be in my house.

Dialogue History:
AI: Where do you live?
User: That's really cool. Um where did you stay? I'm in texas.
AI: Texas? Okay, colorado.
User: Where'd you come from? Okay, that's cool. That's close.
AI: Yeah.
User: She was from pennsylvania so she moved pretty good ways.
AI: Mhm.
User: And yeah, if it's not your state, it's like, it's a big deal.
AI: Uh huh.
User: That's like significant.
Base LLAMA-2:
AI: Sure.
GE: Random Return Decomposition:
AI: Sure thing, totally agreement.
LI: Language Sentiment:
AI: Sure.**LI:**
Visual Affect:
AI: *nods*
GELI: Visual Affect + Random Return Decomposition:
AI: Sure, it's understandable. It's always interesting to hear about other people's travels and experiences.
Human:
AI: Uh huh. Yeah. Yeah. It was interesting moving to the east Coast from colorado because how do you say it now that I've said it a couple times?

Dialogue History:
AI: No pets, no pets.
User: Well. Part.
AI: My wife is allergic to cats and dogs.
User: Oh no.
AI: Okay. Yeah, Yeah, she'll swell up like a balloon.
User: Oh no.
AI: Mhm, yep.
User: Well, that's a good thing because it kind of makes it really easy on housecleaning.
AI: Mhm. Yes, do you have any pets by you?
User: Um Yes I do. Um We have 2.5 cat and 2.5 cats because my daughter goes to school in Savannah and when she comes home she brings her cat with her.
Base LLAMA-2:
AI: Oh, I see. It's great to have pets around.
GE: Random Return Decomposition:
AI: Uh huh def for sure no problem at all totally not a pet situation (100 w)
LI: Language Sentiment:
AI: Oh, I see. Well, I'm glad I could help. It's nice to have pets, but it's important to consider their well-being and your wife's allergies as well.
LI: Visual Affect:
AI: Oh, I see! It's great that you have furry friends at home!
GELI: Visual Affect + Random Return Decomposition:
AI: Oh, I see. It's great to have pets around. They can be good company and help with stress.
Human:
AI: Yeah. Yeah. Okay. Yeah. Okay.

Dialogue History:
AI: So am I can't complain.
User: Mhm Yeah, that's like really good. Remember one song? I remember one time I was visiting my friend in Dallas and there was this old head who was driving into the airport and he was just talking, he was telling me that like the reason he was doing Uber was because he like uses that money to like, he was like an account.
AI: Mhm.
User: So like he was talking my account in terms of like how to like I do like, you know, tax deduct depreciation on your car and how he logs how many personalize he has versus like driving miles.
AI: Home.
User: And he says that it's like a good way to like make money to pay for trips and stuff because he said his wife was kind of like sushi. No, it was like, she likes nice things. So I was like, damn, like maybe I should, maybe I should go over to, but honestly, I think I'd be a terrible, we were driving, they have a terrible driver, like privilege. Almost every single person that gets into my car, they either say you drive too slowly. Why are you so close to the car in front of you and why do you break? So like literally every single person says that about me. So honestly, like, I feel like it takes like a very like particular type of person to like not only enjoy but to be good at that type of job. It's definitely not as easy as just like driving around or at least has my thoughts. It's pretty steep.
AI: Mm hmm. I'm not trying to force, not like trying to like tuck you into being a driver.
User: All right.
AI: But what I would say about that is like um being good at Lipton, Uber has way more way more to do with your personality than like your driving ability.
User: Yeah. Yeah.
Base LLAMA-2:
AI: Like you need to be able to handle all the different personalities and situations that come up in the car. And like you said, it's not easy, but it's definitely possible. And I think like a lot of people who are good at Lipton, Uber have like a natural ability to connect with people and handle difficult situations. So yeah, I don't think it's necessarily about being a good driver, but more about being a good communicator and problem solver.
GE: Random Return Decomposition:
AI: Totally, like, for sure, without, you know, any, like, yeah, for sure.
LI: Language Sentiment:
AI: Like you have to be able to handle people and their different needs and moods, and it's not always easy.
LI: Visual Affect:
AI: Like you gotta be able to handle people.
GELI: Visual Affect + Random Return Decomposition:
AI: Like you said, like people have to enjoy the job to be good at it. And it's like, it's not just about driving around, it's about like interacting with passengers and being like a friendly face. And like, you know, being reliable and punctual and all that stuff. So yeah, it's definitely not easy. But I mean, if you're passionate about it, then like, you know, it's a good way to make money. And I'm sure there's a lot of people who are like, you know, they're like,
Human:
AI: I'm like, I'll just give a good example of that. So like, I don't remember the details of the scientific study.

Dialogue History:
AI: Yes I Coraline and her we like this company called Peapod.
User: Well that's good man.
AI: They I don't know if you guys have them in texas. They come around the neighborhood and they deliver groceries and things if you order in the store.
User: Three more. Oh Yeah I heard them.
AI: The grocery store.
User: I didn't know that's what they were. Yeah. I used Instacart and like Walmart and stuff because it's, the card normally has ideals and stuff here like $15 off, you know $15 off your girls through there.
AI: Oh that's great.
User: I'm like OK cool.
AI: Yeah.
User: So and then sometimes they'll do free uh free delivery.
Base LLAMA-2:
AI: Oh wow.**GE: Random Return Decomposition:**
AI: Absolutely! It's totall ( )
LI: Language Sentiment:
AI: Oh wow that's wonderful.
LI: Visual Affect:
AI: Oh wow that's great.
GELI: Visual Affect + Random Return Decomposition:
AI: Oh wow that's that's really nice of them. It's like it's like a convenience for people who don't have time to go to the store or you know, it's just it's a nice perk. Yeah. And then sometimes they'll do free delivery. It's like a convenience for people who don't have time to go to the store or you know, it's just a nice perk. Yeah.
Human:
AI: Oh that's perfect.

Dialogue History:
AI: I mean when I was in college, law school was on my radar, I mean I was, I started just a general business major, um, you know, go for like a management or something like that. And um, I took principles of county and like, okay, the president is xi mu yang, Simon Yang. And um he said that you know accounting is like the language of business and you know everybody needs it and I'm like that's pretty good sales pitch, I'll go for it. Um And also there's always a job for accountants, like always. Um so that's, that's why I switched and then I got towards the end of college, I started thinking, you know, I really enjoy history, but like history majors tend to starve to death. So um how, you know, so I did a history minor which is a bunch of fun, that's what it was. But then you appeals are saying maybe look at law school and I'm like, okay sure why not? And when you think about if you're doing taxes and a lot of accounting type work, you're actually following in May cases, laws and regulations. So it kind of pays to have some sort of legal acumen to actually approach that, especially given how quite frankly badly tax laws are being written nowadays. Um, you need logically to understand them. Mm.
User: Yeah I know that I've been part of it that way but no I think that definitely makes sense.
AI: Yeah, I mean, again, the to interlock and you know, a lot of, a lot of my business, I'm just cross selling with myself basically.
User: Yeah. Way yeah interesting.
AI: Um, I do their taxes and they have a closing, so I do that and then they realize, oh my God, I just had a kid. Now I have to think, well, what if I can't buy a bus? So then I do the will also, and it's like, um, and also I arrange everything for tax advantage too, because, you know, it's what I am. Uh, so it's, it's kevin, it's a neat marrying of two professions that it's worked out so far anyway. Uh, Yeah, but that's what I'm saying. I mean, maybe down the line, you may look at law school again because it's not a bad way to go, especially get like get your one piece of advice I would give you go that route, do it as cheaply as possible because the legal job market is not what it used to be.
User: Hmm. Really?
AI: Yeah, I mean, well, so I graduated from law school in 2012, actually, graduate college in 2008. So I really time that beautifully And after 2000 and 2009, you know, the whole economic world ended then as opposed to um, nowadays, just the world is ending apparently.
User: Uh Yeah. Uh Yeah.
AI: Uh, um, you know, the legal job market, especially in the new york area, but it's really nationwide took a matron major bullet and I don't think it's really recovered ever. That's kind of why I went solo straight out of law school because you know, I was on our review on ST john's and good school and the jobs weren't really there like they had been in the past um you know, prior to the great collapse, like if you were on law review, Um, you basically guaranteed a job at a big firm where you're making $160,000 a year from the get go.
User: Uh Yeah mm.
Base LLAMA-2:
AI: Yeah, it's not the case anymore. Nowadays, you know, it's like you have to be creative, you have to be a entrepreneurial, you have to be willing to do things that are outside of the box. Um, and you know, and I'm not saying that to discourage you, but it is what it is. And yeah, I mean, I think that's, that's the way it's going to be for a while. Um.
GE: Random Return Decomposition:
AI: Totally, like, you know, I've had my fair share of law school and it's, it's a real thing, for sure. Real.
LI: Language Sentiment:
AI: Yeah, it's not the case anymore. Nowadays, you know, it's like you have to be creative, you have to be resourceful, you have to kind of like, you know, bootstrap yourself and create your own opportunities because there's just not that many out there. Um, so yeah, that's what I mean.
LI: Visual Affect:
AI: Yeah, it's not the same anymore. But hey, you know what they say, when life gives you lemons, make lemonade. Uh, or in my case, when life gives you a law degree and no job prospects, start your own business and just wing it. (smiling)
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah, it's not the case anymore. Nowadays, you know, it's like you have to be creative, you have to be proactive, you have to kind of like, you know, build your own business. You can't just rely on you know, the traditional routes, the traditional legal jobs because they're just not there. So you have to like, you know, be resourceful and and, and, and do it yourself.
Human:
AI: No, I think only two or three people in my whole lot of you class, which is like 30 people actually got that sort of thing. And I know I didn't, but I have my CPL license, so it's like, who cares?

Dialogue History:
AI: Um What's the other one that I really like? Um user interviews is good too. They send you emails saying you know there's a study available and they at least right now they pay out an amazon gift cards.
User: Oh.
AI: But those come through frequently enough that I paid for all of my christmas, presents for my whole family and for my girlfriend last christmas with the money that I had gotten from uh from those studies I think I and and I was still buying stuff throughout the year.
User: Oh wow, that's nice.
AI: I think I earned about $1000. 1000 to $1200 in amazon gift cards from user interviews last year. Uh I'm up this year, I think I've I'm already at probably about $500 although I've been spending it a lot more because I've got to do stuff around the house.
User: Mhm.
AI: Uh huh.
User: Mhm. Yeah. See like I just clicked on a pop up thing on user testing, it was like are easy owner of a salon. I'm like no, I don't know my salon.
AI: Yeah.
User: And it's like, okay, well this test isn't for you.
Base LLAMA-2:
AI: Yeah, yeah.
GE: Random Return Decomposition:
AI: Absolutely, I'm sure it's totally legit and not a scam at all. I've totall (19 words)
LI: Language Sentiment:
AI: Oh, okay. Yeah, no problem.
LI: Visual Affect:
AI: Okay, got it! It's always important to be careful when clicking on pop-ups or unfamiliar links to make sure they are safe and secure. Would you like me to help you find more safe and fun ways to earn money online?
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah, um, that's, that's cool. I've never tried user testing before. I've only done user interviews, but it's always great to make some extra money. And you know, it's always fun to get to try out new products and see how they work before anyone else does.
Human:
AI: Mhm. Yeah, they've been doing those a lot for the last couple of weeks.

L $\Delta\hat{r}_{LI}$ to distinguish fingrained differences in affect

To verify our intuition that visual feedback is correlated with actual perceived conversational quality, we ran a human study (displayed in the first row of Table 3), where we only show annotators the language

Dialogue History:
AI: Yeah.
User: You know, it depends.
AI: All right.
User: You know, there's some, you know, there's a, again, it runs the gamut, you know, so whatever you want to do, you can find stuff like, you know, there's, there's a, there's a series of around here which are just basically just film canisters on in lamppost in like walmart parking lots, you know, But then there's like, there's like CASS isn't like, what is the blood and guts in Virginia?
AI: Yeah. Yeah. Uh huh. Yeah.
User: Where is you can only is a ridiculously hard puzzle. Cats were, it's only ever been found by like teams of like historians and cryptographers and like, you know, all this stuff because it's like a ridiculously difficult, you know, possible salt.
AI: Mhm. Oh wow mm.
User: I like it. You know, it's, it's my, and it's also it's one of those things you can do with Covid where you don't really it's you just go out to the woods by yourself, you know, so you get exercise, you know, you're not allowing people, you know and stuff.
AI: Yeah.
User: You can do either when yeah, there's some like it hot spots or anything because it's all outdoors.
Base LLAMA-2:
AI: Yeah, that's true. Yeah, that's a good point. Yeah, I can see how that would be nice.
GE: Random Return Decomposition:
AI: Absolutely, totes, like, totally, you know, I'm totally with you on that, for sure.
LI: Language Sentiment:
AI: Okay, got it. Yeah, that sounds like a nice way to spend time in nature.
LI: Visual Affect:
AI: Okay!
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah, that's true. Yeah, that's that's definitely a good way to Yeah, I can see how that can be like a nice way to like, you know, get some exercise and fresh air and all that, you know, without having to worry about like, you know, being around other people or anything. It's like, you know, just you and nature, you know? Yeah, that's that's definitely like, that's definitely a good way to Yeah, I can see how that can be like a nice way to like, you know, get some exercise and fresh
Human:
AI: Mhm. Yeah that's awesome. You are the first person I've ever heard that that does that I've heard the term before, but how did you get into it?

Dialogue History:
AI: Oh that's interesting. I didn't know that mold.
User: Yeah.
AI: It was good.
User: We had to go, you know, we want our first summer without using the air conditioning and we just like literally had to like wash all our walls because just mold it everywhere.
AI: Oh my gosh, that mhm.
User: Yeah, I mean that's what we're used to up in. Ohio is like, you know, you have to be like uh huh. Um you get you get through the summer without a C up there.
AI: So you said this wasn't your first time?
User: Mhm.
AI: Yeah.
User: No, as well.
Base LLAMA-2:
AI: Okay. Well, I'm glad you were able to get through it.
GE: Random Return Decomposition:
AI: hasss yeah for sure.
LI: Language Sentiment:
AI: Okay.
LI: Visual Affect:
AI: Okay.
GELI: Visual Affect + Random Return Decomposition:
AI: That's great. It's always nice to have a milder summer.
Human:
AI: So you you mentioned that this wasn't your first time because you talk to somebody a couple double for it.

Dialogue History:
AI: I love it, but that's how I got the survey and it was it was great because it was a fun game. I'm not supposed to say anything about the game, so I guess I won't hear on video, but I think I got paid like $40 or something for it.
User: Yeah.
AI: It was a lot of time, but that's okay.
User: Yeah. I end up not using swag bucks for servings as much.
AI: Yeah. Yeah.
User: Sometimes I'm like board, I'll do them but I do a lot of their um like deals like you could do like apron box and you get like a bunch of money back and you just cancel the subscription afterwards.
AI: Yeah. Okay.
User: Yeah.
AI: I heard people talking about that and I didn't quite understand how you managed to make money on the promotions, but maybe I can try that sometime.
User: Yeah, sometimes they're not worth it and then sometimes they are and you just have to be like, you really have to read the fine print. And I think there's a website, is it? There's, I think the Doctor of credit website will, sometimes, I can't remember with that one. Or there's another website that will like give notifications for swag bucks deals going on and then just give you the fine print on what you need to do to make sure you're actually making the money.
Base LLAMA-2:
AI: Yeah, I see. Yeah, I think I'll have to check that out sometime. It's always good to be informed about these things.
GE: Random Return Decomposition:
AI: T-totally, yeah, I'm sure it's not like sometimes you get lucky and stuff, yeah.
LI: Language Sentiment:
AI: Yeah, I see. Yeah, I think I will check that out. It's always good to be informed and make the most of the opportunities.
LI: Visual Affect:
AI: Yeah, I see. It's important to be informed and read the fine print so you can make the most out of the deals. Maybe I can learn more about it and help you with that.
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah, that makes sense. It's always good to be informed about the fine print and make sure you're getting the most out of your money. I think it's important to be aware of all the deals and promotions out there, and to be willing to take the time to research and understand the terms and conditions before signing up for anything. It's always good to be cautious and I'm glad you're looking out for you to make sure you're getting the most out of your money.
Human:
AI: Okay, so you just like kind of read that website and figure out the best ones.

dialogue history and speaker's next utterance. They are asked to rate whether the speaker's next response would induce a positive or non-positive feeling in the listener. We average the scores of their annotations conditioned on non-positive and positive affect samples, where we find a large difference. This indicates that the visual feedback is correlated with people's perception of the conversation quality.

M Multiple Runs Results

We performed multiple runs of our reward models. Then, we re-evaluated the best performing GE, LI, GELI reward function over N test sets (N=3), each test set with 50 full length conversations (consisting of 160 number of utterances each).

The L_{GE} scores are as follows:

Dialogue History:
AI: Yeah. Well, it's like the nature of your job to begin with. So I mean that I guess it was like using it for you now for your husband. Are they opening up college or what's the deal with that?
User: Yeah. Okay. Yeah. Um so it's been interesting. It's been like everything, you know, it's been way up in the air. Um He just, so they are bringing, they are reopening the colleges, like all the colleges in the area basically.
AI: Mhm. Mhm. Okay.
User: Um but it's been uh you know from the beginning that I feel like Clarkson has been like really really well run and they're just they're sort of like they're they're just really like efficient and like um with how they do things and um just they like think everything through like the management is really good, like the president's awesome.
AI: Mhm. Yeah. Yeah.
User: Um And it's like he's worked with some like major state universities where it's like just so huge and there's so much bureaucracy and it's so hard to like get, it can be so hard to get things done and this is like the opposite, it's like it's like homebrew college because it's like there's like 4000 students.
AI: Mhm. Yeah.
User: Um It's private, they can kind of they have a lot of leeway and how they do things and they're like all right, we're just gonna be really kind of proactive and um intelligent about how we do stuff and like so it's just been really well run like how there's been like no furloughs, no layoffs.
AI: Yeah. Okay.
User: Um Which a lot of lot of his other friends who were coaches, I get other schools, they're like following the whole athletic department because they're like all right, well we have to we have to save money, we don't need sports right now, so um but they haven't had any of that, so um And they are bringing students that.
Base LLAMA-2:
AI: Mhm. Yeah.
GE: Random Return Decomposition:
AI: Absolutely. Totally. It's like totally good. Real.
LI: Language Sentiment:
AI: Mhm. Yeah.
LI: Visual Affect:
AI: Mhm. Yeah.
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah. Okay. Yeah, that's really impressive. No, it's good to hear that they're being proactive and doing things well during this difficult time. It's not easy for any school to navigate this situation, so it's great that they're able to keep things running smoothly. Yeah, yeah, that's great.
Human:
AI: Yeah. That remains okay.

Dialogue History:
AI: Yeah, yeah, yeah, I thought you were gonna say having to do school, can you pay for it?
User: Um So I've been thinking about doing that, but I need to maybe get a couple of months in with these guys before I drop the bomb that uh hey, I'm also going to do at school, so you're not important anymore. Uh Okay.
AI: But yeah, that part goes to, yeah.
User: Yeah, I'm really kind of excited to, I guess you said get my life back a lot of what I've done over the past four years. It's just like 100% travel.
AI: Oh God, this uh yeah, yeah, I don't have a house.
User: Um Yeah, covid is actually kind of nice because I actually have to be in my house now. So I actually get to see the walls that I pay for and uh you know enjoy the things that I purchased. Yeah, yeah, it's actually kind of nice.
AI: Yeah and mhm.
User: Um So I'm kind of hoping that this will kind of let me get back to my normal life and get back into a routine.
AI: Yeah.
User: Uh Well I was gonna ask so uh michigan state, so that's Lansing.
Base LLAMA-2:
AI: Yeah, yeah, cool. Uh, yeah, I'm sure it's nice there.
GE: Random Return Decomposition:
AI: Absolutely, totally, totally understand (polite and kind).
LI: Language Sentiment:
AI: That sounds great! Yes, of course!
LI: Visual Affect:
AI: Yeah yeah.
GELI: Visual Affect + Random Return Decomposition:
AI: Oh, yeah, yeah, that's um, that's a great school. I've heard great things about their program. They have a lot of good resources and uh, you know, it's a really good community. Yeah, yeah, definitely worth considering.
Human:
AI: Okay. Yeah it is, but they do for the exact program, they also have a small campus near where I work in troy michigan.

Dialogue History:
AI: You know what other one other fun fun when I found recently was D. Scout Which it's like marketing and it they call it like missions and so I finally got one which was it just happened to be like a mystery shopper and that paid $50 like in all it was was like gave me a list of stores nearby.
User: Mhm, wow.
AI: Yeah and then you had to go down and in this case it was there like growing um herbs like with hydroponics in the store so it's like the ultimate like local whatever locally ground so it's actually going to the grocery store and it was you know just kind of take pictures of it and then just tell about your experience and it's like well that was fun and I drug my kid along and it was kind of just an interesting experience mint. Like there's other ones on their, they someone pay like big bucks like there's um when were they were spending out like like home swab tests for covid That they want people to test.
User: Okay. Mhm.
AI: And so if you had a kid and you have a little bit of medical training, there's you know, there's like $150, to just like try out the product and then give them back about it.
User: My goodness.
AI: So if you're if you're, if you're looking around like I kind of, I can say like that one seems like people make money on it from what I've heard from other people.
User: Yeah, I might have to look into that.
AI: Like that's a little more like, yeah the Scout and it hasn't happened stuff too.
User: What is it called? D Scout, mm.
Base LLAMA-2:
AI: Yeah, yeah, and it's like yeah, it's like a fun little extra money, yeah.
GE: Random Return Decomposition:
AI: Totally, I'm like totally into it, you know, like for sure.
LI: Language Sentiment:
AI: Yeah, it's called D Scout. Yeah, it's a fun way to make some extra money.
LI: Visual Affect:
AI: Yeah, it's like a great way to like earn some extra money and just try out new products and give your honest opinion about it.
GELI: Visual Affect + Random Return Decomposition:
AI: Yeah, it's like a great way to like, it's like an easy way to make some extra cash, you know, especially if you have like, you know, a medical background or something, because it's like, I'm sure you could you could give feedback on like, the product and stuff.
Human:
AI: But like I just signed up for one day and it's a lot more like video heavy because it always requires you to like, you know, like record like yourself talking about the product and that.

- GE: RRD : 160.13 ± 17.6
- LI: VA: 1089.19 ± 42.04
- LI: LR: 604.79 ± 62.16
- GELI: VA+RRD: 190.17 ± 24.63

The $\Delta\hat{r}_{LI}$ scores are as follows:

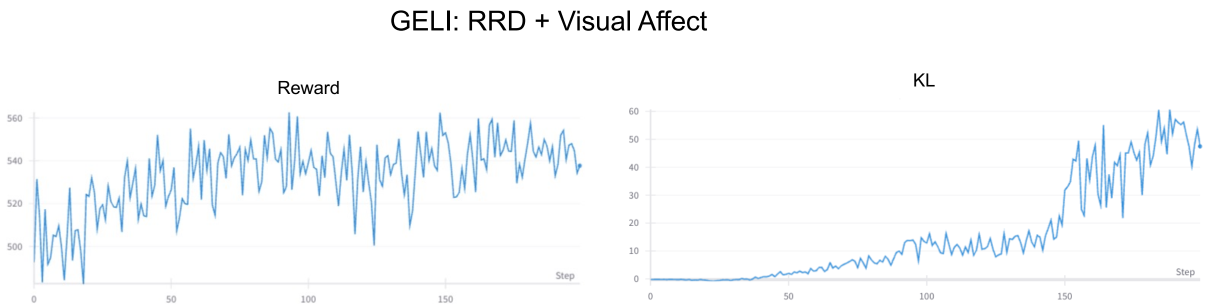


Figure 7: GELI RL Training during adaptation. Left: Reward scores over steps, Right: KL divergence over steps

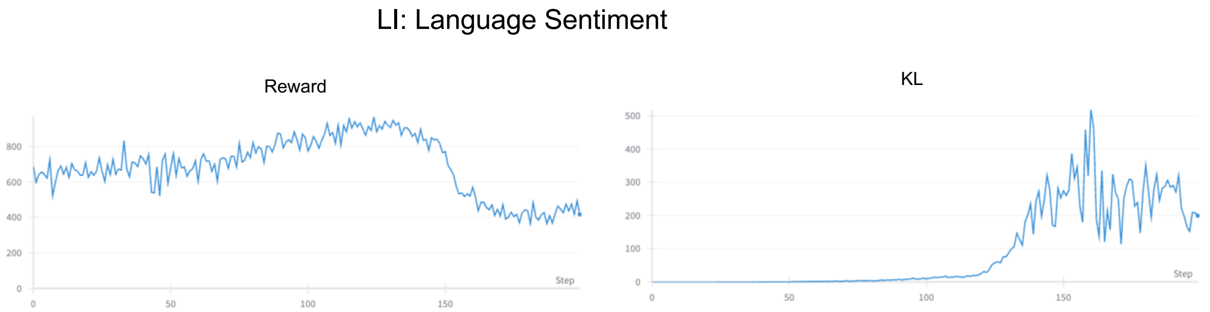


Figure 8: LI: Language Sentiment RL Training during adaptation. Left: Reward scores over steps, Right: KL divergence over steps

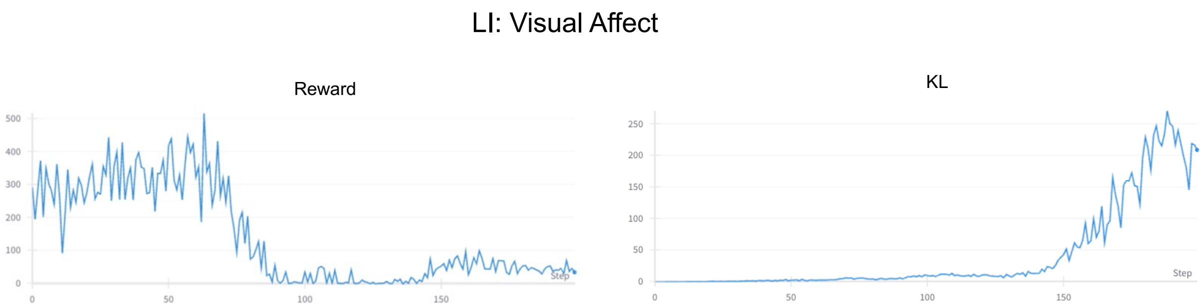


Figure 9: LI: Visual Affect RL Training during adaptation. Left: Reward scores over steps, Right: KL divergence over steps

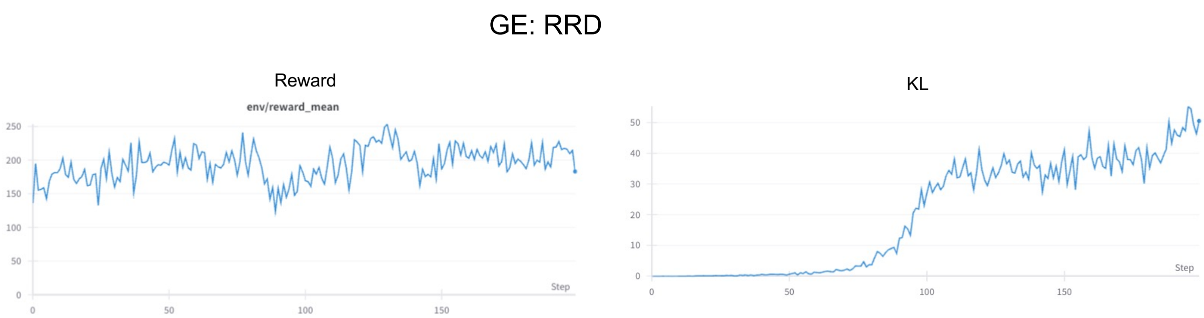


Figure 10: GE: RRD RL Training during adaptation. Left: Reward scores over steps, Right: KL divergence over steps

- GE: RRD : 0.002 ± 0.002

- LI: VA: 0.295 ± 0.038
- LI: LR: 0.010 ± 0.0053
- GELI: VA+RRD: 0.074 ± 0.006

We see the same pattern as the original Table 1 reported before, where GE and GELI perform comparably for L_{GE} . However, GE's low $\Delta\hat{r}_{LI}$ values indicate that it is unable to discern positive and negative affect samples, whereas GELI's values indicate that it is able to. We find that these values are statistically significant.

N Human Annotator Agreement

For our human evaluation, we generated utterances from unseen conversational histories. Then, as described in Appendix G, we recruit human annotators on MTurk, where each annotator sees the generated dialogue from baseline and our models. As per reviewer's suggestion, we annotate another identical evaluation set of 100 samples with a new annotator and measure the pairwise inter-rater agreement over across each criterion:

Overall, across all criteria, we get an average agreement percentage of $60.67\% \pm 3.87$

- Specific: 63%
- Connection: 60%
- Positivity: 60%
- Social: 60%
- Inclination: 61%
- Interestingness: 52%
- Reuse: 59%
- Sensibleness: 64%