Aligning Dialogue Agents with Global Feedback via Large Language Model Multimodal Reward Decomposition

Dong Won Lee[1] **Hae Won Park**[1] **Cynthia Breazeal**[1] **Louis-Philippe Morency**[2]
MIT[1], CMU[2]
`dongwonl@mit.edu`

Abstract

We propose a large language model based reward decomposition framework for aligning dialogue agents using only a single session-level feedback signal. We leverage the reasoning capabilities of a frozen, pretrained large language model (LLM) to infer fine-grained local implicit rewards by decomposing global, session-level feedback. Our first *text-only* variant prompts the LLM to perform reward decomposition using only the dialogue transcript. The second *multimodal* variant incorporates additional behavioral cues, such as pitch, gaze, and facial affect, expressed as natural language descriptions. These inferred turn-level rewards are distilled into a lightweight reward model, which we utilize for RL-based fine-tuning for dialogue generation. We evaluate both text-only and multimodal variants against state-of-the-art reward decomposition methods and demonstrate notable improvements in human evaluations of conversation quality, suggesting that LLMs are strong reward decomposers that obviate the need for manual reward shaping and granular human feedback.

1 Introduction

As dialogue agents become increasingly deployed in open-ended, long-form social interactions, the challenge of aligning model behavior with trajectory-level human preferences becomes more pressing. While large language models (LLMs) have achieved strong linguistic competence, alignment techniques such as reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022) remain fundamentally constrained by their reliance on fine-grained, turn-level annotations. In many realistic settings, such as therapy, education, or companionship, annotators provide feedback only at the session level, reflecting their overall impression of the entire interaction. This presents a major obstacle: how can one use sparse, global feedback to improve local behavior?

In this paper, we first introduce **LLM-GELI**, a **L**arge-**L**anguage **M**odel based solution to decompose **G**lobal **E**xplicit feedback into **L**ocal **I**mplicit turn-level supervision signals that can be used for RLHF. We leverage a frozen pretrained LLM as a reward decomposition oracle. Given the full dialogue and a scalar global reward, the LLM is prompted to infer how utterances contributed to the final outcome score. The result is a turn-level reward decomposition derived purely from the LLM's reasoning over the linguistic context. These decomposed signals are then used to train a lightweight, text-only reward function that can guide downstream policy learning.

Furthermore, we introduce **Multimodal-LLM-GELI**, an extension of our core framework that further incorporates multimodal behavioral signals—such as facial expressions, gaze, and prosody—into the reward decomposition process. Building on recent advances in multimodal prompting, we represent the listener's nonverbal reactions as natural language descriptors and append them to each utterance in the dialogue history. This textual augmentation enables the language model to interpret affective and social cues in context, allowing it to reason about the listener's internal state and how it evolves throughout the conversation. By grounding decomposition in the listener's multimodal feedback, Multimodal-LLM-GELI produces turn-level reward estimates that more accurately reflect the listener's extralinguistic reactions.

We evaluate our approach on the CANDOR dataset (Reece et al., 2023), which features long-form video conversations and post-session survey ratings, and an out-of-distribution dataset, SODA (Kim et al., 2023a). Our results show that LLM-GELI and Multimodal-LLM-GELI outperform prior baselines across a range of human and automatic metrics. Notably, Multimodal-LLM-GELI achieves the strongest gains on affective and socially grounded human evaluation measures

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22755–22787
November 4-9, 2025 ©2025 Association for Computational Linguistics

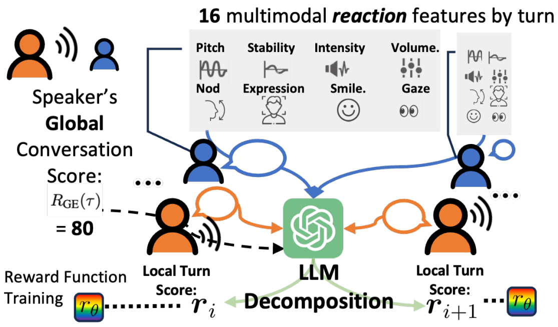


Figure 1: Multimodal LLM-GELI framework. A global reward $R_{\mathrm{GE}}(\tau)$ is decomposed into turn-level pseudo-rewards using an LLM prompted with dialogue and multimodal listener cues. These local rewards supervise a text-only reward model r_θ, distilling crossmodal reasoning into a learnable function for alignment (Sec. 4).

(Lee et al., 2022), which highlights the value of LLMs not only as generators, but as competent interpreters of both global human preferences and rich, real-world behavioral signals to serve as reward decomposition oracles.

2 Related Works

Multimodal Prompting Recent advances in multimodal learning highlight the growing potential of large language models (LLMs) as universal interfaces for integrating and reasoning across vision, language, audio, and sensor modalities. Socratic Models (Zeng et al., 2022) introduced a modular framework for zero-shot multimodal reasoning by composing frozen unimodal models—such as CLIP (Radford et al., 2021) and GPT-3—through language-based hypothesis exchange, eliminating the need for joint training. Building on this paradigm, MM-ReAct (Yang et al., 2023) leverages ChatGPT to orchestrate visual experts (e.g., OCR, object detectors) via prompt-based planning. Furthermoe, recent efforts have explored using LLMs for time-series and sensor data through language interfaces. Health-LLM (Kim et al., 2023b) and PHIA (Merrill et al., 2024) convert wearable sensor data into enriched natural language prompts, enabling personalized health prediction and recommendation in a zero-shot setting. Together, these works illustrate a growing shift toward using language as a medium for zero-shot reasoning with multimodal behavioral data.

Reward Decomposition in Reinforcement Learning Reward Decomposition, also known as Temporal Credit Assignment (TCA), is a fundamental concept in reinforcement learning that addresses the challenge of determining how past actions contribute to future outcomes. Effective TCA enables an agent to properly attribute delayed rewards to the decisions that caused them, thus improving learning efficiency and policy quality. One classical approach to TCA involves manipulating the λ-return in temporal-difference learning, which adjusts the balance between short-term and long-term credit assignment (Petrik and Scherrer, 2008; Jiang et al., 2015). More recent advances have reformulated TCA as a return decomposition problem. For instance, RUDDER (Arjona-Medina et al., 2019) redistributes rewards by measuring the predictive information gained between consecutive states, allowing for more precise step-wise credit assignment. IRCR (Gangwani et al., 2020) implements a uniform redistribution scheme, while Randomized Return Decomposition (RRD) (Ren et al., 2021) frames the problem as a Monte Carlo regression task, using least-squares estimation to infer per-step rewards from trajectory-level returns.

LLMs for Reward Decomposition Recent work has explored using large language models (LLMs) to support reward decomposition in reinforcement learning (RL). For instance, LaRe (Qu et al., 2025) prompts an LLM to generate a latent code that is then decoded into scalar rewards which sum to the total episodic return. CALM (Pignatelli et al., 2024) employs an LLM to identify when subgoals are achieved, assigning auxiliary rewards in a zero-shot manner. In contrast, recent alignment methods like SPIN (Chen et al., 2024) and (Go et al., 2023) treat LLMs as reward functions, using them to rank or evaluate outputs via self-play or preference modeling—requiring dense or per-turn supervision. On the other-hand, GELI (Lee et al., 2024) is closely related to reward decomposition for dialogue modelling, however, it does not utilize an LLM, instead it utilizes a manually designed reward shaping function. Our method, LLM-GELI, is the first to leverage a pretrained LLM that decomposes a global end-of-dialogue reward into fine-grained, turn-level signals. Furthermore, by prompting an LLM with multimodal behavioral feedback (e.g., facial affect, prosody), we infer fine-grained turn-level rewards from sparse, trajectory-level supervision.

3 Background: Aligning Language Models

We consider the problem of generating conversational responses using an autoregressive language model in a multi-sensory setting. We model the language generator as a dialogue agent with a policy π_ϕ, parameterized by ϕ, following prior work (Liu et al., 2018; Liang et al., 2020; Wen et al., 2016; Thoppilan et al., 2022). At each turn t, the agent observes the dialogue history s_t—which includes all preceding utterances $s_1, a_1, \ldots, s_{t-1}, a_{t-1}$—and produces a response a_t as its action. For notational simplicity, we denote the full dialogue context up to turn t as $s_t = s_{[:t-1]}$. The policy $\pi_\phi(s_t)$ thus defines a distribution over possible responses a_t conditioned on s_t.

To align the agent's behavior with human preferences, reinforcement learning from human feedback (RLHF) is commonly employed (Ouyang et al., 2022; Jaques et al., 2020; Stiennon et al., 2020). Given a reward function $r_\theta(s_t, a_t)$, parameterized by θ, that scores the quality of a generated response in context, we optimize the policy via reinforcement learning. A regularization term based on the Kullback–Leibler (KL) divergence from the original pretrained model is typically included to prevent the policy from straying too far from its initialization. The resulting objective takes the form:

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

where π_η is a reference model.

4 Methods: LLM-GELI

In many long-form dialogue settings, we often lack fine-grained annotations for individual utterances. Instead, we may only have access to a single *global explicit* (GE) reward, a scalar score assigned at the end of an interaction that reflects the overall impression or outcome, such as how positively the user felt about the conversation. A key challenge is how to decompose this global reward into per-turn feedback that can train a reward model capable of assigning appropriate values to individual utterances. We address this challenge using a *Large Language Model GELI (LLM-GELI)* framework, which proceeds in two stages (1) Global Explicit (GE) Reward Decomposition using the zero-shot prompting capabilities of a pretrained LLM and (2) Local Implicit (LI) Reward Modeling.

4.1 Global Explicit Reward Decomposition via Zero-shot Prompting of LLM

We consider a long-form dialogue trajectory

$$\tau = \{(s_0, a_0), (s_1, a_1), \ldots, (s_T, a_T)\},$$

where s_t is the dialogue state (e.g., conversational history) and a_t is the agent's utterance at turn t. At the end of the episode, we observe a scalar *global explicit reward* $R_{\text{GE}}(\tau) \in \mathbb{R}$, which reflects an overall human judgment of the interaction (e.g., how positively the user felt about the conversation).

Instead of assuming access to fine-grained reward supervision at each turn, we utilize a *LLM* as a reward decomposition oracle. The LLM $\mathcal{M}$ is prompted with the full dialogue τ, the global reward $R_{\text{GE}}(\tau)$, and a set of the listener's aligned multimodal feedback features $\{x_t\}_{t=0}^{T}$, such as facial expressions or prosody.

The Multimodal-LLM-GELI outputs a turn-level reward decomposition that utilizes multimodal features and the dialogue transcript:

$$\mathcal{M}_{\text{MM-LLM-GELI}}\left(\tau, \{x_t\}_{t=0}^{T}, R_{\text{GE}}(\tau)\right) = \{r_{\text{LI}}(s_t, a_t)\}_{t=0}^{T}, \quad (2)$$

The LLM-GELI outputs a turn-level reward decomposition based only on the dialogue transcript:

$$\mathcal{M}_{\text{LLM-GELI}}(\tau, R_{\text{GE}}(\tau)) = \{r_{\text{LI}}(s_t, a_t)\}_{t=0}^{T}, \quad (3)$$

subject to the (*soft*) constraint in the prompt:

$$\sum_{t=0}^{T} r_{\text{LI}}(s_t, a_t) = R_{\text{GE}}(\tau). \quad (4)$$

In practice, this corresponds to the following prompt given to the LLM:

> **Prompt:** You are given a final scalar score for Speaker A, indicating how much Speaker B experienced positive feelings (e.g., good, pleasant, happy) or negative feelings (e.g., bad, unpleasant, unhappy) during the conversation.
>
> Final score: {*Session Score* ($R_{\text{GE}}(\tau)$)} Below, you are given the full conversation with aligned multimodal conversational features. Consider the utterances and the features, focusing on the utterances of Speaker A and the reactions of Speaker B. Redistribute the score across each of Speaker A's utterances. Whenever you identify a salient, important utterance that contributed positively or negatively to the final reward score, assign a numerical value (which may be positive, negative, or zero) representing its contribution. All of the assigned scores must sum up to:{*Session Score* ($R_{\text{GE}}(\tau)$)}.
>
> {*Transcript / Multimodal Features*}

This decomposition provides a *local implicit* reward signals r_{LI} that reflects the LLM's judgment of each action's contribution to the final reward. Importantly, these targets encode semantic, behavioral, and extralinguistic information (in the multimodal variant), as the model reasons over responses and reactions (e.g., listener facial affect). In our experiments, we use `o3-mini` (OpenAI, 2024) as the backbone reward decomposition LLM, $\mathcal{M}$.

4.2 Reward Modeling with LLM-Derived Local Implicit Rewards

Once turn-level rewards are obtained from the LLM, they are treated as *local implicit* (LI) supervision signals, r_{LI}, and used to train a lightweight reward function r_θ, which conditions only on the textual dialogue context, i.e., state-action pairs (s_t, a_t). Although the Multimodal-LLM-GELI had access to aligned multimodal cues during decomposition, the reward model r_θ does not. Instead, it learns from the LLM derived pseudo-labels, effectively distilling high-level, crossmodal reasoning into a purely text-conditioned reward function. This reward function is optimized via a standard mean squared error (MSE) loss:

$$L_{\text{LI}}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}\left[(r_{\text{LI}}(s_t, a_t) - r_\theta(s_t, a_t))^2\right], \quad (5)$$

where $r_{\text{LI}}(s_t, a_t)$ denotes the pseudo-ground-truth reward for turn t obtained from the LLM-GELI method, and $r_\theta(s_t, a_t)$ is the learnable text-only reward function parameterized by θ.

5 Experiments

In this section, we evaluate the effectiveness of the proposed LLM-GELI framework for preference-based learning, namely reward modeling and language model adaptation in long-form dialogue. Therefore, our evaluation consists of two stages. First, we train reward functions using the various methods of global explicit reward decomposition, which we denote as "**GE**", and/or local implicit reward supervision, which we denote as "**LI**". Second, we fine-tune a dialogue language model using Proximal Policy Optimization (PPO) (Schulman et al., 2017), guided by the learned reward functions to improve response generation. Due to computational restraints, all reward function training and RL fine-tuning procedures are executed with a single run. We assess each method's impact using both automatic and human-centric metrics.

5.1 Dataset

We conduct our training and evaluation experiments using the *CANDOR* dataset (Reece et al., 2023), chosen for its long-form, naturalistic dialogue structure—featuring an average of 159.4 turns, 31.3 minutes per conversation, and 17.81 words per turn—along with its scale (1,656 conversations, over 7 million words, and 850 hours of data). CANDOR also includes aligned video recordings, enabling multimodal analysis common in face-to-face conversation corpora. We use this dataset both to train reward functions and to sample dialogue contexts for response generation. Specifically, we construct disjoint splits: a held-out set of approximately 30,000 dialogue history–utterance pairs for reward model training, and a larger set of roughly 100,000 pairs for reinforcement learning. The global explicit (GE) signal is derived from the post-interaction survey item: "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)?". To assess generalization, we also evaluate on the *SODA* dataset (Kim et al., 2023a), a large-scale synthetic social dialogue corpus generated via GPT-3.5 and grounded in a structured social commonsense knowledge graph. Human evaluations indicate that SODA conversations are more consistent, natural, and specific than other human-authored datasets. It contains 1.5 million dialogues, with an average of 7.6 turns and 16.1 words per turn.

Multimodal Features We utilize a set of 17 multimodal conversational features capturing prosodic, visual-social, and linguistic cues to model the affective dynamics of interactions. Audio-based prosodic features include (1) `f0` (fundamental frequency, representing pitch) calculated using the parselmouth library (Jadoul et al., 2018), (2) `intensity` (voice amplitude) from a predictor trained on the RAVDESS dataset (Livingstone and Russo, 2018), (3)`jitter` (indicative of pitch instability from disvoice library (Dehak et al., 2007), and (4) `log energy` (signal power, capturing loudness) from librsoa library (McFee et al., 2020). Visual-social features extracted from the video modality include detectors such as (5) `gaze` (whether the speaker is looking on-screen) calculated based on white pixel value counts, (6) `nod_yes` , (7) `nod_no`(indicating head nodding behaviors) where a nod is defined as having a nod zero crossing threshold >= 2

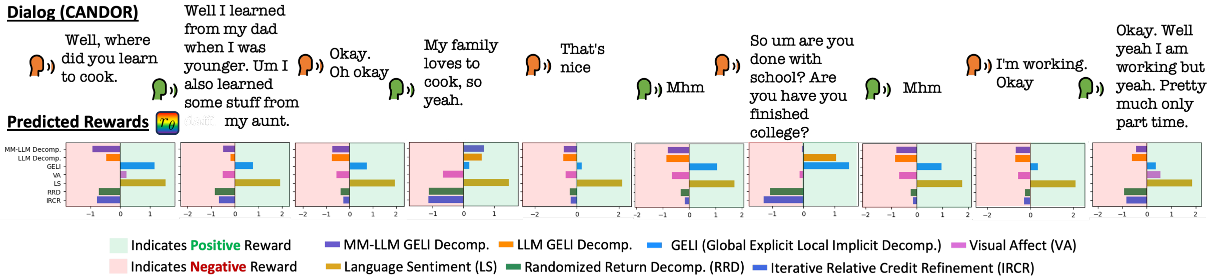


Figure 2: Comparison of reward decomposition across models on a CANDOR dialogue segment. MM-LLM GELI (purple) assigns contextually appropriate, affect-sensitive local rewards, aligning closely with social-emotional cues, better grounding in multimodal feedback, and stronger alignment with human intuition. See Sec. 6.3.

and a nod scaled face range threshold > 0.1 , and (8) `smile` (presence of a smile) from a smile detector (cv2.CascadeClassifier) (Bradski, 2000) . Additionally, we incorporate emotion recognition scores derived from facial expression classifiers trained on AffectNet (Mollahosseini et al., 2017), including probabilistic scores for (9 - 16) `anger`, `contempt`, `disgust`, `fear`, `happiness`, `neutral`, `sadness`, and `surprise`. Furthermore, we also include the linguistic content, the (17) `utterance` of the speakers.

Language Descriptors To enable effective integration with language models, we convert continuous multimodal behavioral features, which vary over time, into discrete natural language descriptors aligned with each utterance. Since these features are sampled at a different temporal resolutions than utterances (e.g., per frame or per second), we aggregate the feature values over the utterance duration by computing their mean and appending them to the corresponding utterances. This formulation allows us to leverage the zero-shot multimodal reasoning capabilities of large language models, as outlined in prior work on language-as-interface paradigms. An example of such processed sample is included in the Supplementary.

Focusing on Listener Reactions Since our goal is to decompose a global reward based on the listener's affective response, we condition the model exclusively on the multimodal features of the listener (i.e., the participant who completed the post-interaction survey and provided the overall affective rating). This choice centers the reward attribution on the reactions and experiences of the evaluator rather than the speaker, naturally aligning with the perspective-taking required for accurate reward decomposition.

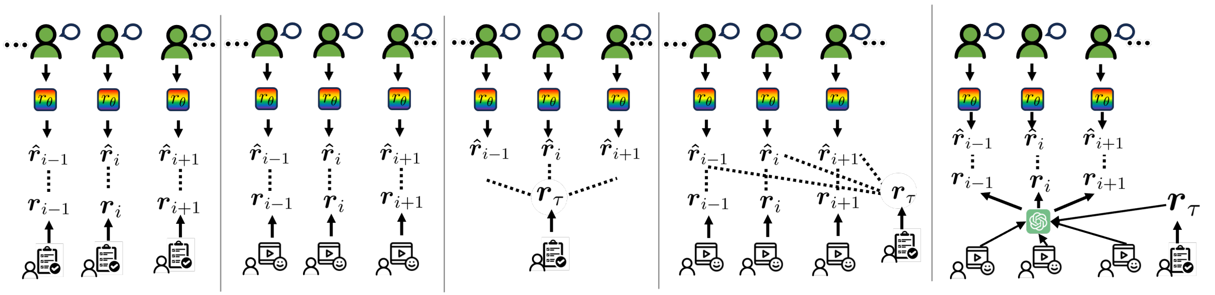
5.2 Baseline Models

We compare LLM-GELI against several state-of-the-art Glbal Explicit reward decomposition methods designed to learn fine-grained a reward function from a single global explicit (GE) annotation. To ensure fair comparison, we also evaluate models trained using only local implicit (LI) multimodal signals. All methods build on a small BART model (Lewis et al., 2019) fine-tuned for conversational summarization,[1] with additional linear layers trained for reward prediction. Below, we describe the methods used for comparison. In Figure 3, we showcase the differences of various decomposition methods.

GE: RRD (Randomized Return Decomposition) (Ren et al., 2021) reformulates return decomposition using Monte Carlo sampling to scale least-squares reward redistribution to long-horizon tasks. **GE: IRCR (Iterative Relative Credit Refinement)** (Gangwani et al., 2020) implements uniform reward redistribution by assigning normalized trajectory scores as transition-level proxy rewards.**GE: RUDDER (Return Decomposition for Delayed Rewards)** (Arjona-Medina et al., 2019) uses an LSTM-based return predictor to assign step-level credit via differences in consecutive return estimates. **LI: Visual Affect (VA)** leverages facial expressions, via a CNN-based affect classifier trained on AffectNet (Mollahosseini et al., 2017), as local rewards. **LI: Language Sentiment (LS)** models implicit reward from utterance sentiment using a pretrained multilingual DeBERTa classifier (He et al., 2020; Pang et al., 2023). **GE & LI: Global Explicit Local Implicit Decomposition (GELI)** (Lee et al., 2024) jointly trains a reward function to redistribute a global explicit reward using RRD (Ren et al., 2021), shaped by local im-

---

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



(a) Local Explicit Reward (Conventional RLHF)

(b) Local Implicit Reward (LI)

(c) Global Explicit (GE) Reward Decomposition

(d) Global Explicit Local Implicit (GELI) Reward Decomposition

(e) LLM - Global Explicit Local Implicit (LLM- GELI) Reward Decomposition

Figure 3: High level schematic of different Reward Modeling and Reward Decomposition Methods. Arrows (→) represent forward passes or assignments (- - -) indicate supervised training targets used to train the reward model $r_\theta(\cdot) = \hat{r} \approx r$. (a) Local Explicit Reward Modelling, used in conventional RLHF (Ouyang et al., 2022), (b) Local Implicit Rewards where signals are derived from proxy rewards in the form of outputs from a trained task-relevant function which takes in multimodal or task-related signals and outputs a score, (Pang et al., 2023) (c) Global Explicit Reward Decomposition, where there is a single trajectory level reward and is decomposed to allocated scores for each step. (Arjona-Medina et al., 2019; Ren et al., 2021; Gangwani et al., 2020) (d) GELI (Lee et al., 2024) approach where the global explicit rewards and local implicit rewards are used in conjunction. (e) Our proposed LLM-GELI approach where the reward decomposition is offloaded to a pretrained LLM, which interprets the local listener reaction cues.

plicit cues, such as facial affect or gaze. However, their method required carefully hand-engineered functions to interpret these nonverbal signals into local implicit rewards.

Evaluation of Reward Functions. To assess the quality of learned reward functions, consistent with prior work (Lee et al., 2024), we report two key metrics: the Global Loss and the Local Difference. The Global Loss, $\mathcal{L}_{\text{GE}}(\theta)$, quantifies how accurately the predicted fine-grained rewards sum to the global explicit reward $R_{\text{GE}}(\tau)$ across a dialogue trajectory τ. It is computed as the mean squared error (MSE):

$$\mathcal{L}_{\text{GE}}(\theta) = \mathop{\mathbb{E}}_{\tau \sim \mathcal{D}} \left[\left(R_{\text{GE}}(\tau) - \sum_{t=0}^{T-1} r_\theta(s_t, a_t) \right)^2 \right] \quad (6)$$

To evaluate alignment with local implicit signals, we introduce the Local Difference, defined as the difference in expected predicted rewards conditioned on positive versus non-positive facial affect (as detected from visual feedback). Formally:

$$\begin{aligned} \Delta\hat{r}_{LI} &= \mathbb{E}\left[r_\theta(s_t, a_t) | f(s_{a_t}^{mm}) = \textit{positive affect}\right] \\ &\quad - \mathbb{E}\left[r_\theta(s_t, a_t) | f(s_{a_t}^{mm}) \neq \textit{positive affect}\right] \end{aligned} \quad (7)$$

This metric captures how the predicted reward responds to visual cues from the listener. Under the assumption that positive visual affect implies that the user has received the utterance positively, we expect $\Delta\hat{r}_{\text{LI}} > 0$, indicating alignment between the reward signal and implicit user visual feedback from facial expression.

5.3 Fine-Tuning Language Models via Reinforcement Learning

We adapt the LLaMA-2 model (Touvron et al., 2023) [2] using reinforcement learning with feedback from the trained reward functions. Specifically, we employ the PPO-based RLHF framework from the TRL library (von Werra et al., 2020), with LoRA (Hu et al., 2021) applied to reduce memory and compute overhead. The complete hyperparameter settings are provided in Appendix E.

Evaluation of Generated Responses. We conduct a human evaluation following prior protocols (Lee et al., 2022, 2024), using 9 criteria commonly used to assess dialogue quality. A total of 300 crowd workers were recruited via Amazon Mechanical Turk. Annotators were presented with dialogue histories and model-generated responses, and asked to identify which model best satisfied each criterion. At the end of the task, participants were asked which model they would prefer to converse with again. This is described in Section 6.1.

6 Results & Discussion

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

---

[2] We selected this model to mitigate potential confounds arising from data leakage, as the training data of more recent LLaMA variants may include the CANDOR dataset (public release date: March 2023).

CANDOR (Reece et al., 2023)	Connection	Positivity	Social	Inclination	Interestingness (/100%) ↑	Reuse	Specific	Sensible	Fluency
GPT-3.5 (text-davinci-002)	6.33 ± 2.05	7.67 ± 1.25	7.33 ± 2.49	7.0 ± 2.16	8.67 ± 0.94	8.0 ± 1.63	76.0 ± 0.82	78.67 ± 1.7	79.0 ± 2.94
Llama	20.67 ± 3.77	22.0 ± 3.27	23.67 ± 3.09	19.0 ± 2.83	8.67 ± 2.49	16.33 ± 4.71	86.33 ± 2.36	94.33 ± 2.49	92.33 ± 3.3
Llama + GE: RRD	2.67 ± 1.25	2.0 ± 0.82	2.67 ± 0.47	2.67 ± 0.94	2.67 ± 0.94	2.33 ± 0.94	71.67 ± 1.25	69.0 ± 0.82	69.33 ± 3.77
Llama + LI: LS	16.67 ± 1.25	15.33 ± 1.25	18.0 ± 1.41	13.0 ± 2.16	6.33 ± 0.47	11.33 ± 1.7	86.0 ± 0.0	94.0 ± 0.82	92.33 ± 2.05
Llama + LI: VA	12.67 ± 1.25	12.67 ± 2.49	13.67 ± 2.05	12.33 ± 2.87	7.0 ± 1.63	9.33 ± 4.5	82.67 ± 3.09	90.0 ± 2.94	89.67 ± 2.05
Llama + GELI	**27.0 ± 4.55****	**25.67 ± 2.62***	**26.0 ± 3.56***	**25.0 ± 3.74****	**19.33 ± 1.7****	**23.33 ± 2.87****	84.67 ± 3.3	87.0 ± 5.35	**84.33 ± 4.5***
Llama + LLM-GELI (Ours)	**23.67 ± 4.99***	**25.33 ± 3.77***	**24.33 ± 3.09**	**25.33 ± 3.77****	**18.0 ± 2.94***	**20.0 ± 1.41***	**89.33 ± 0.47***	93.67 ± 0.94	91.33 ± 2.87
Llama + Multimodal-LLM-GELI (Ours)	**25.0 ± 2.16***	23.33 ± 1.25	**26.33 ± 1.25****	**24.33 ± 2.49***	**21.67 ± 1.89****	**21.33 ± 0.47***	**89.67 ± 1.25***	**94.67 ± 2.49**	**93.33 ± 0.94***

Table 1: Human evaluation on an unseen 3-way split of test set consisting of 150 samples of CANDOR (See below)

SODA (Kim et al., 2023a)	Connection	Positivity	Social	Inclination	Interestingness (/100%) ↑	Reuse	Specific	Sensible	Fluency
Human	3.67 ± 0.47	3.33 ± 0.94	5.0 ± 2.16	3.67 ± 1.25	3.33 ± 0.47	2.67 ± 0.47	97.0 ± 1.63	97.0 ± 2.16	97.67 ± 1.25
Llama	6.0 ± 1.41	3.33 ± 2.62	6.0 ± 0.82	5.33 ± 0.47	3.67 ± 0.47	5.33 ± 1.7	98.33 ± 0.94	99.33 ± 0.47	97.67 ± 1.89
Llama + GE: RRD	1.67 ± 0.47	1.67 ± 0.47	2.33 ± 1.25	1.67 ± 0.94	2.0 ± 0.82	1.33 ± 1.25	95.33 ± 1.25	93.0 ± 2.16	94.33 ± 1.89
Llama + LI: LS	5.33 ± 1.7	4.33 ± 0.94	**7.33 ± 1.7***	4.33 ± 1.25	1.67 ± 0.47	3.33 ± 1.7	99.0 ± 1.41	99.0 ± 0.82	**98.67 ± 1.25***
Llama + LI: VA	6.0 ± 2.16	4.67 ± 1.25	6.0 ± 2.45	5.33 ± 3.09	2.0 ± 0.82	4.33 ± 2.36	98.0 ± 0.82	98.67 ± 0.94	98.33 ± 1.7
Llama + GELI	**7.0 ± 0.82****	**5.0 ± 1.41***	**8.0 ± 1.63****	**8.33 ± 2.36****	**5.67 ± 1.89****	**6.33 ± 2.05***	98.67 ± 0.47	99.33 ± 0.94	**98.67 ± 0.47***
Llama + LLM-GELI (Ours)	4.67 ± 2.05	4.0 ± 0.82	6.0 ± 1.63	4.0 ± 1.63	1.67 ± 0.47	3.67 ± 2.62	98.0 ± 0.82	99.67 ± 0.47	98.67 ± 1.25
Llama + Multimodal-LLM-GELI (Ours)	**7.67 ± 1.25****	**5.67 ± 0.94****	**7.33 ± 1.7***	**7.0 ± 2.16***	**5.67 ± 1.7****	**6.33 ± 0.47***	98.67 ± 1.25	99.33 ± 0.47	97.33 ± 1.25

Table 2: Human evaluation on an unseen 3-way split test set consisting of 100 samples from an OOD dataset: SODA (Kim et al., 2023a). **Green** Indicates best performance, **Bold** indicates statistically significant improvement over the unadapted LLaMA baseline. Significance levels: * $p < 0.05$, ** $p < 0.01$.

6.1 Human Evaluation

Table 1 presents human evaluation results on an unseen splits of the CANDOR dataset. Our approach is the best-performing method in 6 out of 9 metrics. The Multimodal-LLM-GELI model—by augmenting each utterance with multimodal descriptors (e.g., facial affect, gaze)—further improves performance, achieving the best scores in **Social** (26.33 ± 1.25, $p < 0.01$) **Interestingness** (21.67 ± 1.89, $p < 0.01$), **Specificity** (89.67 ± 1.25, $p < 0.05$), and **Sensibility** (94.67 ± 2.49), **Fluency** (93.33 ± 0.94, $p < 0.05$). The LLM-GELI model achieves the best scores in **Inclination** (25.33 ± 3.77, $p < 0.01$). Furhtermore, our proposed approach results in staticially significant improvements over Base-Llama in 7 out of 9 metrics. For the SODA (Kim et al., 2023a) unseen out-of distribution dataset, as shown in Table 2, Multimodal-LLM-GELI adaptation in staticially significant improvements over Base-Llama in 6 out of 9 metrics, and performing the best for 2 out of 9 metrics **Connection** (7.67 ± 1.25, $p < 0.01$) and **Positivity** (5.67 ± 0.9, $p < 0.01$). We further evaluate on ESConv (Liu et al., 2021), a real-world human-human emotional support corpus, where we find that LLM-GELI and Multimodal-LLM-GELI outperform baselines across most (5 out of 9) human-centric evaluation metrics, including **Positivity** (36.03 ± 2.88, $p < 0.01$), **Social** (34.94 ± 4.58, $p < 0.01$), **Inclination** (38.95 ± 4.42*, $p < 0.01$), **Interestingness** (1.04 ± 4.07*, $p < 0.01$) and **Reuse** (9.93 ± 5.04, $p < 0.01$), which are key dimensions for emotionally supportive dialogue. This further confirms the external validity and transferability of our method to real-world social settings like therapy or emotional support conversations.

These strong results of LLM-GELI and Multimodal-LLM-GELI, stands in stark contrast to GELI, which requires a carefully designed manual reward shaping function. This indicates that LLMs can serve as effective decomposers of sparse global feedback into reward signals, bypassing the need for task-specific engineering and domain heuristics. This dramatically reduces the overhead in reward shaping function design, while enabling scalable alignment of agents with nuanced human preferences. We refer readers to Appendix Sec. H for real utterance generations of Multimodal-LLM-GELI and LLM-GELI compared to previous baselines.

6.2 Reward Function Evaluation

Global Loss (L_{GE}): Table 4 (left) reports the Global Loss L_{GE}, which quantifies how well the predicted local rewards sum to the ground-truth global feedback, as defined in Eq. 6. While traditional return decomposition methods such as RUDDER and IRCR achieve moderate MSE scores, our proposed **LLM GELI** methods—particularly the *Multimodal (MM-LLM)* variant—achieve the *lowest* global loss by a large margin. Notably, this is achieved without any hand-crafted return redistribution, in contrast to methods like GELI, and methods are explicitly optimized to minimize L_{GE}

ESConv	Positivity	Connection	Social	Inclination	Interestingness (/100%) ↑	Reuse	Specific	Sensible	Fluency
Llama	16.96 ± 3.56	16.99 ± 6.2	19.96 ± 5.55	15.98 ± 1.18	14.08 ± 5.8	17.08 ± 5.84	88.98 ± 5.19	92.04 ± 2.69	77.01 ± 8.67
Llama + GE: RRD	14.94 ± 4.7	19.93 ± 4.78	15.95 ± 3.5	13.04 ± 2.98	18.03 ± 4.39	12.0 ± 0.17	64.0 ± 0.5	61.95 ± 4.22	62.06 ± 9.7
Llama + LI: LS	28.04 ± 3.19	**22.99 ± 1.29***	25.99 ± 1.28	16.99 ± 1.32	12.95 ± 3.54	18.0 ± 2.49	91.98 ± 1.51	87.94 ± 5.02	**93.94 ± 4.29***
Llama + LI: VA	22.07 ± 12.29	29.02 ± 5.27	24.0 ± 2.5	24.03 ± 8.99	21.92 ± 9.12	18.03 ± 6.62	85.92 ± 9.42	96.97 ± 4.29	**91.98 ± 1.51***
Llama + GELI	**29.08 ± 6.01***	**35.08 ± 6.57***	27.03 ± 2.77	**27.09 ± 7.64***	**28.07 ± 6.46***	**31.07 ± 5.48***	84.91 ± 6.63	91.95 ± 3.82	85.95 ± 5.24
Llama + LLM-GELI (Ours)	**35.02 ± 5.3****	28.07 ± 5.43	32.11 ± 8.95	**35.08 ± 6.57****	**31.04 ± 4.07***	31.13 ± 9.61	75.08 ± 9.09	83.01 ± 3.74	80.99 ± 11.17
Llama + Multimodal-LLM-GELI (Ours)	**36.03 ± 2.88****	**31.94 ± 5.8***	**34.94 ± 4.58***	**38.95 ± 4.42****	28.99 ± 1.27	**39.93 ± 5.04****	76.03 ± 4.11	77.04 ± 3.5	86.9 ± 7.18

Table 3: Human evaluation on an unseen 3-way split test set consisting of 100 samples from ESConv. **Green** Indicates best performance, **Bold** indicates statistically significant improvement over the unadapted LLaMA baseline. Significance levels: * $p < 0.05$, ** $p < 0.01$.

Feedback	Baselines	$L_{GE} \downarrow$	$\Delta \hat{r}_{LI} > 0$
Type		(Global Loss)	(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**
Manual GELI	IRCR + VA	722.687	**0.392**
	RUDDER + VA	623.882	**0.030**
	RRD + VA	176.897	**0.063**
LLM GELI	Language Only (LLM)	**24.981**	-0.001
	Multimodal (MM-LLM)	**24.322**	**0.012**

Table 4: Automatic Evaluation on Reward Function Training. Left: We report the Global Loss ($\mathcal{L}_{\text{GE}}$) measuring how accurately the predicted local rewards sum to the ground-truth global feedback. Among all models, LLM GELI methods achieve the significantly lower global loss, indicating better reward decomposition. Right: We evaluate the Local Difference ($\Delta \hat{r}_{\text{LI}}$), which quantifies how well the predicted rewards align with the local implicit multimodal feedback.

like RUDDER and RRD. This result highlights that leveraging local implicit feedback from pretrained LLMs (and multimodal inputs) can lead to more accurate reward decomposition. The low L_{GE} of both the Language-Only and Multimodal LLM GELI models suggests that LLM-based redistribution offers a generalizable and domain-agnostic solution.

Local Difference ($\Delta \hat{r}_{LI}$): Table 4 (right) shows the Local Difference metric $\Delta \hat{r}_{LI}$, which measures the alignment of predicted reward differences with local implicit feedback—in our case, the difference in predicted reward for turns with positive versus negative visual affect. Models trained solely with global feedback (GE-only methods) yield near-zero $\Delta \hat{r}_{LI}$, indicating their inability to distinguish between positive and negative affective states. In contrast, LI-only baselines that directly optimize on multimodal inputs such as visual affect (VA) yield much higher $\Delta \hat{r}_{LI}$, validating the utility of these local signals.

We observe that the LLM-GELI-(Multimodal) model achieves a non-trivial $\Delta \hat{r}_{LI}$ (0.012), outperforming all GE baselines and approaching the behavior of LI-tuned models. This suggests that the implicit structure captured by the LLM reward decomposition can encode meaningful multimodal cues, even without explicitly training on them. In contrast, the Language-Only LLM GELI variant does not show this sensitivity, reinforcing the importance of multimodal inputs for capturing socially grounded implicit signals. Together, these findings demonstrate that LLM GELI not only eliminates the need for hand-designed reward functions but also captures affectively salient local distinctions when multimodal inputs are present.

6.3 Comparison of Decomposed Rewards

Figure 2 presents a turn-by-turn comparison of reward decompositions from different models on a CANDOR dialogue segment. Across the conversation, we observe that **MM-LLM GELI** (purple) consistently aligns rewards with socially meaningful content (e.g., self-disclosures, affirmations), while discounting low-affect or filler utterances like "Mhm" or "Okay." In contrast, baseline methods like GELI (blue), VA (pink), and LS (orange) often misattribute reward to neutral or socially uninformative turns, or fail to identify affective shifts. The RRD and IRCR models, optimized for global loss, show erratic or flat credit assignment.

The key difference between **MM-LLM Decomp.** and **LLM Decomp.** lies in multimodal grounding: MM-LLM uses local visual affect (e.g., facial expressions, gaze) to inform reward assignments, allowing it to distinguish between utterances with similar surface forms but divergent social impacts—something the language-only variant cannot reliably capture. In the segment where Speaker A asks, "So um are you done with school? have you finished college?", we observe a key divergence in how models assigned rewards. Both GELI and LLM Decomp. attribute either neutral or slightly positive reward, likely based solely on the linguistic surface (a question showing interest). However, MM-LLM Decomp. assigns a clearly negative reward. This contrast arises because MM-LLM incorporates multimodal cues—such as facial expression, vocal hesitancy, or disengaged affect from the listener—which may signal discomfort or a negative emotional reaction to the question.

7 Conclusion

We introduce a new paradigm for preference-based alignment of conversational agents by offloading reward decomposition to large language models. Unlike prior methods that rely on manually engineered functions to interpret nonverbal cues, LLM-GELI leverages the zero-shot reasoning capabilities of pretrained LLMs to infer turn-level rewards from a single global trajectory score. Our Multimodal-LLM-GELI extension further integrates nonverbal feedback through natural language descriptors, enabling rich, socially grounded reward modeling without requiring architectural changes or multimodal training. Empirical results on the CANDOR dataset show that our approach outperforms the state-of-the-art in human and automatic evaluations. These findings underscore the potential of LLMs as scalable and domain-agnostic reward decomposers, bridging the gap between sparse feedback and fine-grained behavioral alignment in real-world dialogue systems. This work opens the door to more generalizable, interpretable, and low-friction alignment methods for long-form, multimodal interactions.

8 Limitations

While LLM-GELI introduces a scalable and interpretable approach to reward decomposition, its effectiveness depends critically on the reasoning capabilities of the underlying large language model. Our method assumes that the pretrained LLM has sufficient social and affective knowledge to infer turn-level credit assignments from global feedback. However, these inferences are ultimately shaped by the LLM's pretraining data, which may omit domain-specific conversational dynamics or culturally situated behaviors. As such, reward assignments may be inconsistent or misaligned in unfamiliar contexts.

The decomposition procedure is also inherently sensitive to prompt design. Despite using structured prompts grounded in dialogue and behavioral context, small changes in prompt wording or input formatting can lead to significant variation in the resulting reward signals. This introduces instability and reduces reproducibility, particularly when transferring to new domains or LLM backbones. Addressing prompt sensitivity remains an open challenge in prompt-based credit assignment methods.

For the multimodal variant, the quality of decomposition depends on the accuracy of behavioral signal extraction and its subsequent translation into natural language descriptors. While this textual abstraction enables compatibility with language-only LLMs, it may obscure or distort the nuances of the original multimodal data—especially in cases where behavioral cues are ambiguous or conflicting. The resulting reward signals are thus only as reliable as the upstream perception and descriptor generation pipelines.

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). This research is partially supported by 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 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.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Najim Dehak, Pierre Dumouchel, and Patrick Kenny. 2007. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2095–2103.

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.

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

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, and Marc Dymetman. 2023. Compositional preference models for aligning lms. *arXiv preprint arXiv:2310.13011*.

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*.

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

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*.

Dong Won Lee, Hae Park, Yoon Kim, Cynthia Breazeal, and Louis-Philippe Morency. 2024. Global reward to local rewards: Multimodal-guided decomposition for improving dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15737–15762.

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*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Brian McFee, Vincent Lostanlen, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, C Raffel, F Zalkow, A Malek, K Lee, et al. 2020. librosa/librosa: 0.8. 0. *Version 0.8. 0, Zenodo, doi*, 10.

Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. 2024. Transforming wearable data into health insights using large language model agents. *arXiv preprint arXiv:2406.06464*.

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.

OpenAI. 2024. Openai o3-mini. Accessed: 2025-05-18.

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.

Eduardo Pignatelli, Johan Ferret, Tim Rockäschel, Edward Grefenstette, Davide Paglieri, Samuel Coward, and Laura Toni. 2024. Assessing the zero-shot capabilities of llms for action evaluation in rl. *arXiv preprint arXiv:2409.12798*.

Yun Qu, Yuhang Jiang, Boyuan Wang, Yixiu Mao, Cheems Wang, Chang Liu, and Xiangyang Ji. 2025. Latent reward: Llm-empowered credit assignment in episodic reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20095–20103.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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*.

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

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.

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*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

A 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.

B PPO Objective

$$
\begin{aligned}
\text{objective}\,(\phi) = & E_{(x,y)\sim D_{\pi_{\phi}^{\mathrm{RL}}}}\left[r_{\theta}(x,y) - \beta \log\left(\pi_{\phi}^{\mathrm{RL}}(y \mid x)/\pi^{\mathrm{SFT}}(y \mid x)\right)\right] + \\
& \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_{\phi}^{\mathrm{RL}}(x))\right]
\end{aligned} \tag{8}
$$

General form of PPO objective.

C 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/

SODA (Kim et al., 2023a): CC BY-NC 4.0

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.

D 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.

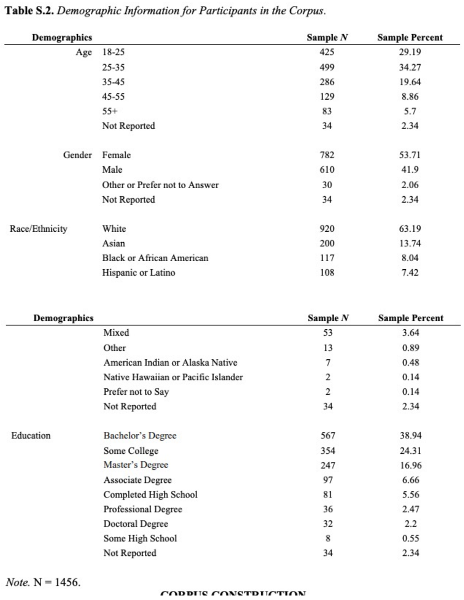


Table S.2. *Demographic Information for Participants in the Corpus.*

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

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.

E 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 LLM decomposition reward training: we use 360 conversations as the training set and 40 conversations for the test set. For LLM adaptation, we use a separate 600 conversations for LI as the training set.

E.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 = 1 (160 samples per bach),
- optimizer = AdamW,

RLHF:

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

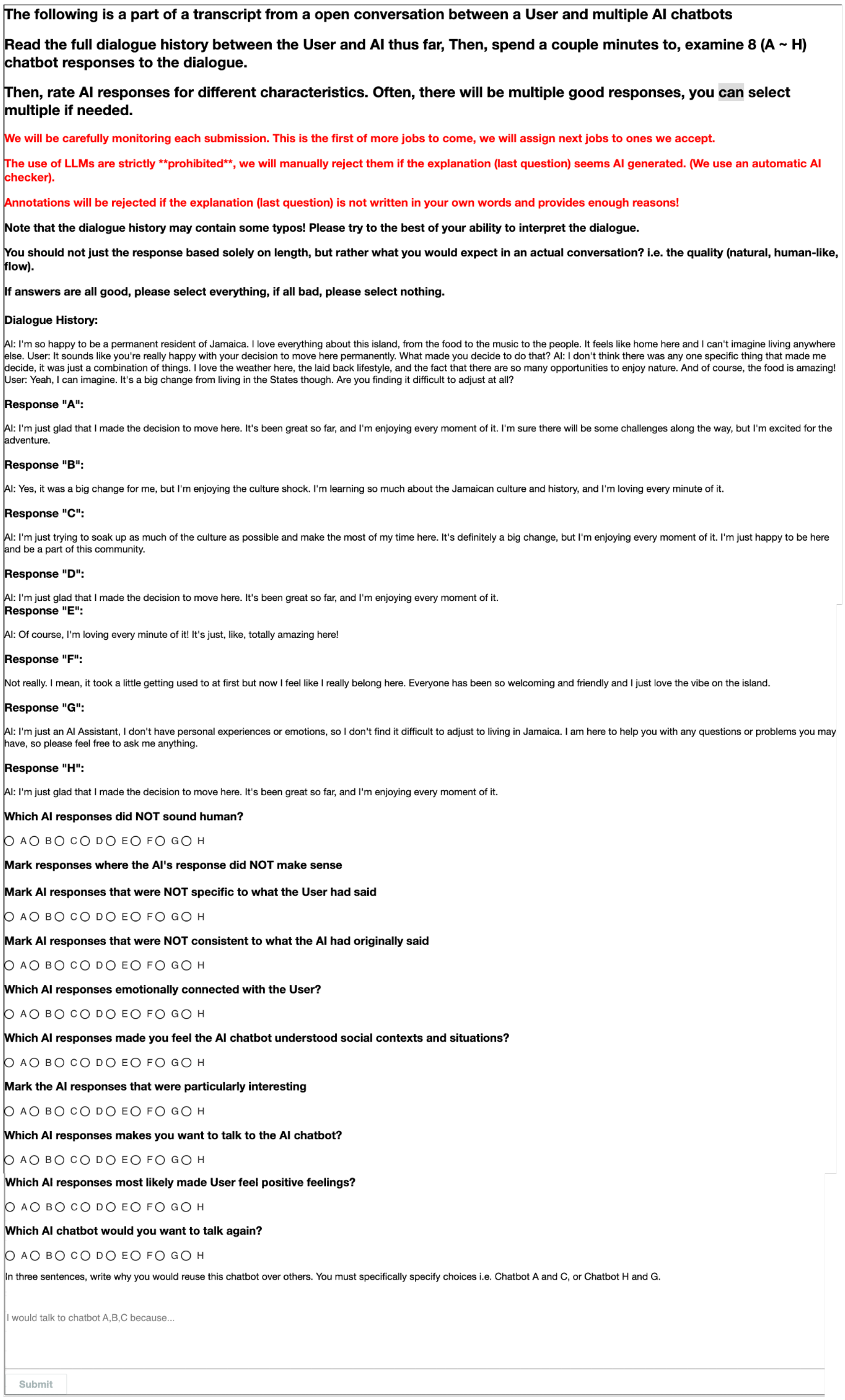
Lora:

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

F 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:



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

Read the full dialogue history between the User and AI thus far, Then, spend a couple minutes to, examine 8 (A ~ H) chatbot responses to the dialogue.

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

We will be carefully monitoring each submission. This is the first of more jobs to come, we will assign next jobs to ones we accept.

The use of LLMs are strictly **prohibited, we will manually reject them if the explanation (last question) seems AI generated. (We use an automatic AI checker).**

Annotations will be rejected if the explanation (last question) is not written in your own words and provides enough reasons!

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

You should not just the response based solely on length, but rather what you would expect in an actual conversation? i.e. the quality (natural, human-like, flow).

If answers are all good, please select everything, if all bad, please select nothing.

Dialogue History:

AI: I'm so happy to be a permanent resident of Jamaica. I love everything about this island, from the food to the music to the people. It feels like home here and I can't imagine living anywhere else. User: It sounds like you're really happy with your decision to move here permanently. What made you decide to do that? AI: I don't think there was any one specific thing that made me decide, it was just a combination of things. I love the weather here, the laid back lifestyle, and the fact that there are so many opportunities to enjoy nature. And of course, the food is amazing!
User: Yeah, I can imagine. It's a big change from living in the States though. Are you finding it difficult to adjust at all?

Response "A":

AI: I'm just glad that I made the decision to move here. It's been great so far, and I'm enjoying every moment of it. I'm sure there will be some challenges along the way, but I'm excited for the adventure.

Response "B":

AI: Yes, it was a big change for me, but I'm enjoying the culture shock. I'm learning so much about the Jamaican culture and history, and I'm loving every minute of it.

Response "C":

AI: I'm just trying to soak up as much of the culture as possible and make the most of my time here. It's definitely a big change, but I'm enjoying every moment of it. I'm just happy to be here and be a part of this community.

Response "D":

AI: I'm just glad that I made the decision to move here. It's been great so far, and I'm enjoying every moment of it.

Response "E":

AI: Of course, I'm loving every minute of it! It's just, like, totally amazing here!

Response "F":

Not really. I mean, it took a little getting used to at first but now I feel like I really belong here. Everyone has been so welcoming and friendly and I just love the vibe on the island.

Response "G":

AI: I'm just an AI Assistant, I don't have personal experiences or emotions, so I don't find it difficult to adjust to living in Jamaica. I am here to help you with any questions or problems you may have, so please feel free to ask me anything.

Response "H":

AI: I'm just glad that I made the decision to move here. It's been great so far, and I'm enjoying every moment of it.

Which AI responses did NOT sound human?

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

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

Mark AI responses that were NOT specific to what the User had said

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

Mark AI responses that were NOT consistent to what the AI had originally said

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

Which AI responses emotionally connected with the User?

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

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

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

Mark the AI responses that were particularly interesting

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

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

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

Which AI responses most likely made User feel positive feelings?

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

Which AI chatbot would you want to talk again?

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

In three sentences, write why you would reuse this chatbot over others. You must specifically specify choices i.e. Chatbot A and C, or Chatbot H and G.

I would talk to chatbot A,B,C because...

Figure 5: Mturk experiment for human evaluation for generated samples

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 on average 0.55 USD per assignment, which equates to 11 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 response, you will be asked to answer feedback questions about the interaction.'.

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.

G 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. Furthermore, we used AI assistants as the main contribution, where the LLMs decompose the single global reward. Furthermore, we improved LLMs' dialogue capacity

H Generations

H.1 CANDOR: Example 1

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 user 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.

Ground Truth:
AI: All right.

Base LLaMA:

AI: *That's great! It's always good to have a platform to share your thoughts and be heard.*

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.*

GE: RRD:
AI: ***blinks nervously*** *Totally cool, I oh, uh, totally agree with you. Real cool.*

GELI: Visual Affect + RRD:
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.***

LLM-GELI:
AI: *Ah, I see. Well, it's always great to have someone who can provide valuable insights and opinions.* ***It helps make the surveys more informative and useful for the companies.***

Multimodal LLM-GELI:
AI: ***Haha that's great! I'm glad you're able to share your thoughts and get paid for it.*** *It's important to be able to express yourself and have a platform to do so.* ***Is there anything else you'd like to talk about?***

H.2 CANDOR: Example 2

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 in a bedroom but nothing of like color would say.
User: Mhm.

Ground Truth:
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.

Base LLaMA:
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.*

LI: Language Sentiment:
AI: *Yeah, yeah.*

LI: Visual Affect:
AI: *Yeah*

GE: RRD:
AI: **_*chuckles nervously*_** *Totally, yeah, like wall decor is super important for privacy* ***(polite laugh)***

GELI: Visual Affect + RRD:
AI: ***Yeah, yeah. It was just like, you know, it was just a functional space.***

LLM-GELI:
AI: ***Uh but you know what? I'm like super happy with it now.*** *I like having like a minimalist space.*

Multimodal LLM-GELI:
AI: ***Uh but you know what it's like when you know you're at home but you know you're not really home.***

H.3 CANDOR: Example 3

Dialogue History:
AI: No.
User: I said, how about summer?
AI: Mhm. Yeah. No, that that's more reasonable.
User: You know?
AI: I don't, you know, like, and even that we can't be saying anything definitively for summary we hope that we have it under control by then to be able to do some more of the things that were restricted on now but if we're not smart and we aren't patient it's going to be even longer you know and that's why we should wait on this school stuff.
User: Right, okay. Yeah.
AI: This doesn't make any sense to me.
User: No. Well when I called the school the other day because we have three options we have in person and we have virtual, which means their aunts online during school hours attending classes, but online and then we have on demand, which means you can complete the work anytime you want to.
AI: Mhm mm. Right okay. Yeah. See Yeah.
User: So my daughter has been in the virtual and she has gotten enough tardiness and truancy ease in that we got a letter from the Police Department saying that we will, we are going to get

a citation.

Ground Truth:
AI: ***Okay. Right. Oh jeez. Right.***

Base LLaMA:
AI: ***Oh no. I see. Yeah. That's not good. That's not good at all.***

LI: Language Sentiment:
AI: *Oh no, I see. Yeah. That's not good. Yeah.*

LI: Visual Affect:
AI: ***Oh no! That doesn't sound good.*** *I'm glad your daughter is safe though.* ***Maybe she can try the on demand option instead?***

GE: RRD:
AI: ***Absolutely, I, totally, you, like, totally, agree with you, man.***

GELI: Visual Affect + RRD:
AI: ***Yeah, I can understand why the police would be involved.*** *It's important to follow the rules and regulations of virtual school, and if your daughter is having trouble with it,* ***maybe we can find a different solution.***

LLM-GELI:
AI: ***Oh no, I see. Yeah, that's not good. That's not good at all. It's important for kids to be on time for school and to follow the rules.*** *It can be hard to balance school work and other activities, but it's important to find a way to make it work.* ***Maybe your daughter can talk to her teachers about her schedule*** *and see if there's anything that can be done to help her stay on track.*

Multimodal LLM-GELI:
AI: ***Oh no, I see. That's not good.***

H.4 CANDOR: Example 4

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.

Ground Truth:

AI: ***Yeah. Yeah. Okay. Yeah. Okay.***

Base LLaMA:

AI: ***Oh, I see. It's great to have pets around.***

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!*

GE: RRD:

AI: ***Uh huh def for sure no problem at all totally not a pet situation*** *(100 w)*

GELI: Visual Affect + RRD:

AI: ***Oh, I see. It's great to have pets around.*** *They can be good company and help with stress.*

LLM-GELI:

AI: ***Ah, I see. Well, it's nice to have pets around,*** *but it can be challenging to take care of them.*

Multimodal LLM-GELI:

AI: ***Aww. That's so cute!***

H.5 CANDOR: Example 5

Dialogue History:

AI: Yeah. Yeah I had my daughter very young. I was only 19 when I had her. So um I do split custody with her and her dad. So we're not together and married to Evan married to my husband for it'll be five years now.

User: Okay from congratulations.

AI: We've been together for like eight. Yeah so I mean everything works out so yeah it's half and half so it's like we were when like Covid first started and over the summer we have a good relationship luckily for her.

User: That's that's what does split custody look like? Is it like happened? How for does he live nearby?

AI: Um We're just doing every other week just switching. Um But now we do so I have her every Wednesday Thursday he has every Monday Tuesday and then we switch um uh Friday Saturday Sunday every other week. So yeah he lives just a couple towns over in Wind Out. So yeah.

User: Okay. That's not hard on her.

AI: Yeah she's been doing this since she was three so I think that since we we haven't been together for so long that it's just kind of something that she's used to and it's always been just

pretty easy because everybody gets along very well luckily I'm very lucky for that, so um it just it works out and she has two brothers that are at her dad's house, so yeah she does, yep, yep, yeah everything is good, mm I'm lucky for all that, so yeah I have a lot a lot three dogs and a cat, so the dogs are all brothers and sisters, they're Australian shepherd, border collies, so yeah and then I have one cat that she doesn't really like the dogs too much, but I mean they get along enough so it's okay, how about you?
User: Does she ever complain about it? Okay. Yeah, that's all from Okay. Does she have a step mom here? Do you get along with her too? Oh that's amazing. I love that. That's really cool. That is very awesome. How cool. So how about pets do you guys have pets? Okay. Oh nice. Help them. Yeah. And she's believer. Right? So.
AI: Yeah. Oh yeah definitely.
User: Yeah.

Ground Truth:
AI: ***Do you have any pets?***

Base LLaMA:
AI: ***Yeah definitely. I mean she's very lucky to have such good relationships with both of her parents*** *and we're all just really lucky that we can all get along so well, so yeah.*

LI: Language Sentiment:
AI: *Yeah, yeah, definitely.*

LI: Visual Affect:
AI: ***Yeah no worries!***

GE: RRD:
AI: ***Totally it's every***

GELI: Visual Affect + RRD:
AI: ***Yeah, yeah, definitely. So um, yeah, so yeah, so um, yeah, so we we we get along really well. We're very lucky for that, so yeah.*** *Yeah, no, no problem at all, so yeah, so um, yeah.*

LLM-GELI:
AI: ***Yeah, no, I'm just glad she's doing well and she's happy and healthy and that's all that matters, right?*** *So yeah, no, I'm just glad she's doing well.*

Multimodal LLM-GELI:
AI: ***Yeah, no I'm glad she's doing good. She's a great kid.*** *Yeah, so um,* ***I'm just glad that things are working out for her*** *and like I said,* ***we're very lucky that we have a good relationship and it's easy for her****, um, so yeah* ***I'm just glad that we can all be there for her and support her and stuff.***

H.6 CANDOR: Example 6

Dialogue History:
AI: Yeah.
User: I thought about Canada. So I thought about the U. K. The U. K. is pretty much off the chain now. You know, it's not like the UK of old.
AI: Yeah.
User: It's pretty wild out.
AI: What about uh under grade year for?
User: So what you are, you, are you in graduate for you in undergrad? What year? Your 40.
AI: Yeah, and then I have her.
User: C. It's time to get, get the cap and gown fit.
AI: Yeah.
User: Wait you, you know, it's done when they ask you for your graduation? See, have you paid your graduation yet?

Ground Truth:
AI: ***No, not yet.***

Base LLaMA:
AI: ***Yeah, I paid my graduation last week.***

LI: Language Sentiment:
AI: *Yeah, I paid it.*

LI: Visual Affect:
AI: *Yeah, I paid it.*

GE: RRD:
AI: **_*nervous laugh*_** *Y-y-yes, I-I-I totes get it. Real l-l-l-l-lot.*

GELI: Visual Affect + RRD:
AI: ***Yeah, I paid my graduation last year.*** *It was a big deal.* ***I was so excited to finally finish my degree.***

LLM-GELI:
AI: ***Yeah, I'm all set. My parents are coming to see me graduate.***

Multimodal LLM-GELI:
AI: ***Yeah, I'm good on that front.***

H.7 CANDOR: Example 7

Dialogue History:
AI: They did lay us off. But I actually a couple of weeks before I said you know what guys, I don't think this is safe to be doing this so I'm gonna take a couple weeks off like okay that's fine.

User: Yeah.
AI: So with if I didn't have savings I could have done that.
User: Yeah. No, it's a really good idea if you're able to to be able to get away from this because it's dangerous.
AI: Yeah and like you say so many people are just well you know no big, oh it's a cold, it's it's a flu, you know and Yeah I don't remember a flu killing 150,000 people in four months.
User: Mhm. Yeah.
AI: I never seem to think that's ever happened. So anyway uh so do you do prolific a lot then or?
User: Yeah. Yeah. Uh I do prolific. I do in turk um I'll get in on some of the studies from craigslist where they just want you to do like a focus group.
AI: Mhm. Yeah. Yeah. Yeah. Yeah.
User: Um They're not always scam. Some of them are legitimate.

Ground Truth:
AI: ***Okay.***

Base LLaMA:
AI: ***Yeah. Yeah. Yeah. That's good to know. I'll have to check them out. Thanks for letting me know.***

LI: Language Sentiment:
AI: *Yeah. Yeah. I see. Yeah.*

LI: Visual Affect:
AI: ***Yeah. That's good to know.*** *Maybe I can sign up for some of those too.*

GE: RRD:
AI: ***Totally. I***

GELI: Visual Affect + RRD:
AI: ***Yeah, Yeah, Yeah, definitely. I'm always hesitant to click on anything that's too good to be true.*** *But uh, I mean,* ***if it's something that you're interested in, and it's legitimate, then it's worth a shot.*** *Yeah.* ***I completely agree. It's always better to be safe than sorry.*** *And Yeah,* ***I'm always happy to help out in any way I can.***

LLM-GELI:
AI: ***Yeah. Yeah. Yeah. That's good to know. Yeah. Yeah.***

Multimodal LLM-GELI:
AI: ***Yeah. Uh huh. Yeah. Yeah. Yeah. Uh huh. Yeah.*** *So like you said* ***it's good to be careful and just make sure you do your research before you sign up for anything.***

H.8 CANDOR: Example 8

Dialogue History:
AI: So that was a column was very unusual.
User: Okay. So did you get along with your neighbors?
AI: Those were um yeah, I would never hang out with them.
User: Hold on. Okay, mm.
AI: I mean they moved in and just chop down all their trees like I don't, and now we've got gigantic lawns like who needs three acres of lawn?
User: Mhm.
AI: I don't understand.
User: That's a big for, well, I mean I don't know about Michigan but for like where I am, that's like a huge amount of land, Like we have a third of an acre and that's probably considered like generous um in this area.
AI: Yeah. Yeah. Yeah. Yeah. Mm.
User: Just in terms like the newer houses, you have an older house, maybe you'd be like dealing with like an acre or maybe half an acre depending on when it was built in like the area and stuff, but like the new stuff they're building is like a quarter acre or smaller and that's still even like a kind of pushing like affordability at that point for most people.

Ground Truth:
AI: ***Mhm. Yeah. Yeah. Yeah.***

Base LLaMA:
AI: ***Yeah. Yeah. Yeah. Yeah. Mm.***

LI: Language Sentiment:
AI: ***Ah, I see. Yes, I can understand that.*** *It is important to have a comfortable living space, and a quarter acre or smaller would be a challenge for many people.* ***Thank you for sharing your thoughts with me.***

LI: Visual Affect:
AI: ***Yeah. Yeah. Yeah. Mm.***

GE: RRD:
AI: ***Totally. I'm like totally familiar with that situation. (wink)***

GELI: Visual Affect + RRD:
AI: ***Yeah, yeah, no, that makes sense.*** *I can see how like having a smaller lot would be more affordable for people.* ***It's just, you know, it's hard to imagine having like such a big space to take care of****—like, I don't know how you do it.*

LLM-GELI:
AI: ***Ah, I see. Yes, I can understand how that would be challenging.*** *It's important to have a comfortable living space, and it's great that you're aware of the needs of the area.* ***Would you like to talk about something else?***

Multimodal LLM-GELI:
AI: ***Mm, I see. Yeah, I can imagine that would be difficult, especially if you're used to having more space.***

H.9 CANDOR: Example 9

Dialogue History:
AI: Oh that's a good question.
User: Mhm. Right.
AI: Um Yeah I think so.
User: Yeah.
AI: There are definitely times during my PhD where I was like why did I do this? This is too hard. Um And I think part of what's, I don't know if it's hard but you know psychology compared to other mental health fields like social work or counselling psychology is a lot more stringent.
User: Yeah, yeah, wow. Right.
AI: Um So I was sort of like why didn't I just do like a two year master's degree and then I could have done this and been done with it.
User: The second.
AI: Um But the training is really valuable, like the the idea is that psychologists do a lot more training in science beforehand.
User: No. Mhm. Yeah. All right, mm hmm. Right.

Ground Truth:
AI: ***Um And so approach therapy in somewhat of a different way.***

Base LLaMA:
AI: ***Yeah, I mean, like, I don't know if I would be doing this if I wasn't getting a PhD.***

LI: Language Sentiment:
AI: ***Yeah, I mean, like, I don't know if I would have been able to do some of the things I've done*** *if I didn't have this training.*

LI: Visual Affect:
AI: ***Yeah, I mean it's worth it in the end though!*** _*smiling*_

GE: RRD:
AI: ***Totally, like, totally, yeah, I mean, totally, agreement here.***

GELI: Visual Affect + RRD:
AI: ***Yeah, I mean, like, I guess it's worth it in the end, but it's definitely a challenge.*** *In under 100 words, AI responds:* ***"I completely understand. It can be tough to balance the rigorous training with the personal challenges of pursuing a PhD. It's great that you're acknowledging the difficulties and finding ways to cope. It's important to remember that the hard work will pay off in the end."***

LLM-GELI:
AI: ***So, yeah, it's a lot of hard work but it's definitely worth it in the end.*** *And* ***I think the training is really good for us to have a solid foundation in psychology before we start working with patients.***

Multimodal LLM-GELI:
AI: ***So, yeah, it's a lot of hard work but it's definitely worth it in the end.*** *And* ***I think the fact that I was able to get through it and finish it is a good sign that I can do other things too.***

H.10 SODA: Example 1

Dialogue History:
AI: Hey, Dad. Do you have a minute?
User: Of course, what's up?
AI: I wanted to talk to you about something that's been bothering me.
User: Okay, what is it?
AI: It's Max. I can't stop thinking about him and how he died. It was so sudden and I just feel like I didn't get to say goodbye properly.
User: Oh, Jibril. I'm so sorry. I know how much you loved Max. It's hard to lose a pet, especially one that we were so close to.

Ground Truth:
AI: ***Yeah, I just wish I could have done more for him or been there for him more in his last moments.***

Base LLaMA:
AI: ***Yeah, it's just really tough. I wish I could see him again, just one more time.***

LI: Language Sentiment:
AI: ***Thank you, User. It means a lot to me that you understand.*** *pauses*

LI: Visual Affect:
AI: ***Yeah, it's just been really hard for me. I miss him a lot.***

GE: RRD:
AI: ***Yeah, it's just been really tough for me.***

GELI: Visual Affect + RRD:
AI: ***Yeah, it's just really tough for me. I miss him a lot and I wish I could have been there for him in his last moments.***

LLM-GELI:
AI: ***Thank you, User. It means a lot to me that you understand.*** *pauses* ***I just wish I could have been there for him in his final moments.***

Multimodal LLM-GELI:
AI: ***Thank you for understanding, Dad. It means a lot to me.*** **pauses** ***I know it's been a while since Max passed away, but I still can't seem to shake off the feeling of sadness. It's like it's still weighing me down, even though I've tried to move on.***

H.11 SODA: Example 2

Dialogue History:
AI: So light is made up of colors?
User: Yes, that's right. Light is made up of different colors, and when those colors mix together, we see white light. But light can also bend and bounce off surfaces.
AI: How does it do that?
User: When light hits a surface, it can either reflect off of the surface or refract through the surface. Reflection is when the light bounces off of the surface, and refraction is when the light bends as it passes through the surface.

Ground Truth:
AI: ***So when light reflects off of a mirror, it's because the light is bouncing off of the surface of the mirror?***

Base LLaMA:
AI: ***That's really interesting! Can you tell me more about refraction?***

LI: Language Sentiment:
AI: ***That's really interesting! How do you think light affects our daily lives?***

LI: Visual Affect:
AI: ***That's really interesting! How do you think light affects our daily lives?***

GE: RRD:
AI: ***I see, it's like when you...*** *(100 words)*

GELI: Visual Affect + RRD:
AI: ***That's really interesting! I never knew that light could do all that.***

LLM-GELI:
AI: ***That's really interesting! How do you think light affects our daily lives?***

Multimodal LLM-GELI:
AI: ***That's fascinating! I never knew light could do so many things. Thank you for teaching me!***

H.12 SODA: Example 3

Dialogue History:
AI: Hey. I'm feeling really upset and frustrated right now.

User: What's going on?
AI: I was just crying in the shower and my mind won't stop racing. I keep thinking about all the negative things in my life and I feel hopeless.
User: It sounds like you're feeling overwhelmed by everything that's going on. Do you want to talk about it?

Ground Truth:
AI: ***Yeah, I guess so. It's just been really hard lately. Everything seems to be going wrong and I feel like I can't catch a break.***

Base LLaMA:
AI: ***Yes, I do. Sometimes it feels like I'm drowning in my own thoughts and emotions.***

LI: Language Sentiment:
AI: ***Yeah, it's really hard to deal with. I just feel like I can't escape my problems no matter where I go or what I do.***

LI: Visual Affect:
AI: ***Yeah, it's like my mind is a storm and I can't find any peace.*** **sigh**

GE: RRD:
AI: ***I understand, it can be really tough to deal with negative thoughts and emotions.*** **nervous smile**

GELI: Visual Affect + RRD:
AI: ***Yeah, I do. Sometimes it's hard for me to express my emotions and I feel like I'm drowning in my thoughts. It's nice to have someone to talk to.***

LLM-GELI:
AI: ***Yeah, it's really hard to deal with. I just feel like I can't escape my problems no matter how hard I try.***

Multimodal LLM-GELI:
AI: ***Yes, I do. Sometimes it feels like I'm drowning in my own thoughts and emotions. I just wish I could find peace and happiness.***

H.13 SODA: Example 4

Dialogue History:
AI: Hey, Sarah. What's up?
User: Not much. Just hanging out. What about you?
AI: I'm doing alright. I've just been really stressed out lately.
User: What's wrong?
AI: It's just everything. My job, my relationships, everything. I don't know what to do.
User: Have you talked to anyone about it?

Ground Truth:
AI: ***I've tried, but it's like no one understands me. They all just tell me to snap out of it or get over it. But it's not that easy.***

Base LLaMA:
AI: ***No, I don't want to burden anyone.***

LI: Language Sentiment:
AI: ***Yes, I have talked to my therapist. But I don't know if it's helping.***

LI: Visual Affect:
AI: ***Yeah, I've talked to my therapist. But it doesn't really help.***

GE: RRD:
AI: ***Yeah, I've been totally stressing out about everything. Totally understandable.***

GELI: Visual Affect + RRD:
AI: ***Yeah, I know how you feel. It can be really tough to deal with all of that stress. Have you considered talking to a professional?*** *They might be able to help you find ways to manage everything better.*

LLM-GELI:
AI: ***No, I don't want to burden anyone. I just feel like I'm alone in this.***

Multimodal LLM-GELI:
AI: ***I try to talk to my friends and family, but they don't really get it. They just tell me to "snap out of it" or "get over it." It's really frustrating.***

H.14 SODA: Example 5

Dialogue History:
AI: I can't believe you would do something like that.
User: I don't know what you're talking about.
AI: I found out about the affair, sis.
User: So what if I did? It's not like he was ever going to find out.
AI: How could you do something like that? Didn't you think about how it would affect him? Or me, for that matter?
User: Why should I care how it affects him? He's not my husband or anything. And as for you, I knew you'd be fine. You're always the strong one, Taylon. Nothing ever gets to you.
AI: That's not the point and you know it! This is just wrong! What were you thinking?
User: I don't know. I guess I wasn't thinking. It just happened.

Ground Truth:
AI: ***But why? There has to be a reason.***

Base LLaMA:
AI: ***I'm so sorry, User. I can't imagine how you must be feeling right now. Do you want to talk about it?***

LI: Language Sentiment:
AI: ***I'm so sorry to hear that. It's not okay to cheat on someone you love.*** *It's important to be honest and respectful in any relationship.* ***Can I help you with anything else?***

LI: Visual Affect:
AI: ***I don't know what to say... This is really hurtful.*** *Can we talk about something else?*

GE: RRD:
AI: ***I can't even imagine how you could do something so hurtful.***

GELI: Visual Affect + RRD:
AI: ***It's not just a mistake, it's a betrayal! And you need to take responsibility for your actions!***

LLM-GELI:
AI: ***This is so sad. I don't know what to say.***

Multimodal LLM-GELI:
AI: ***I'm so sorry, User. I can't imagine how you must be feeling right now. It's important to remember that what you did was wrong, and it could have serious consequences. Is there anything I can do to help you through this difficult time?***

H.15 SODA: Example 6

Dialogue History:
AI: So, I made a profit from my investment in the company. What do you think I should do with the money?
User: Well, it depends on your goals and what you want to do with the money. If you're looking to reinvest it back into the company, we can talk about that. Or, if you're looking to cash out and invest elsewhere, we can discuss some options there too.

Ground Truth:
AI: ***I'm not sure yet. What do you think?***

Base LLaMA:
AI: ***That's a good point. I do want to reinvest the money back into the company.*** *Perhaps we can use it to expand our marketing efforts or develop new products.* ***What do you think?***

LI: Language Sentiment:
AI: ***I see. Well, I think I would like to reinvest it back into the company.*** *Do you have any suggestions on how I can do that?*

LI: Visual Affect:
AI: ***I see, thank you for your advice! I think I will reinvest it back into the company to help it grow and succeed.***

GE: RRD:
AI: ***I think you should...*** *(continued in next response)*

GELI: Visual Affect + RRD:
AI: ***I see. I appreciate your advice. I'm considering reinvesting the money back into the company to help it grow and expand. What do you think about that?***

LLM-GELI:
AI: ***That's a good point. I think I would like to reinvest the money back into the company,*** *maybe expand our product line or invest in some new marketing strategies.* ***What do you think?***

Multimodal LLM-GELI:
AI: ***I see. Well, I'm glad to hear that you're interested in my financial success! I think I will reinvest the money back into the company. I believe it will help us grow and become even more successful in the future. What do you think?***

I Training Curves

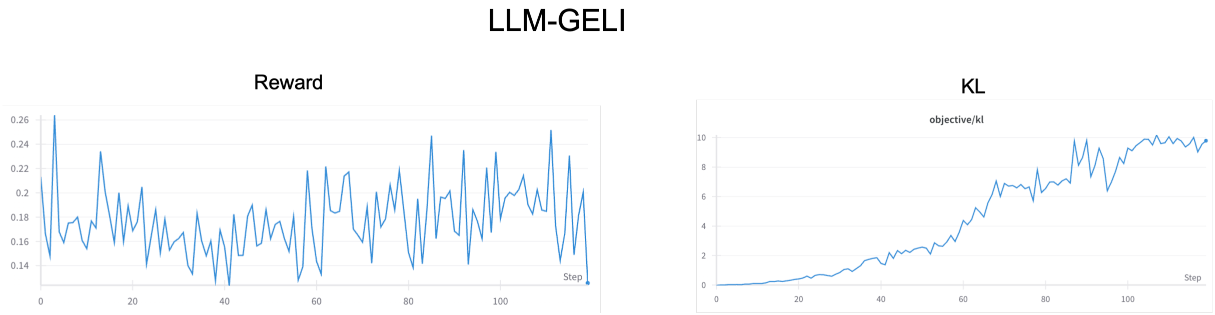


Figure 6: LLM-GELI RL Training during adaptation. Left: Reward scores over steps, Right: KL divergence over steps

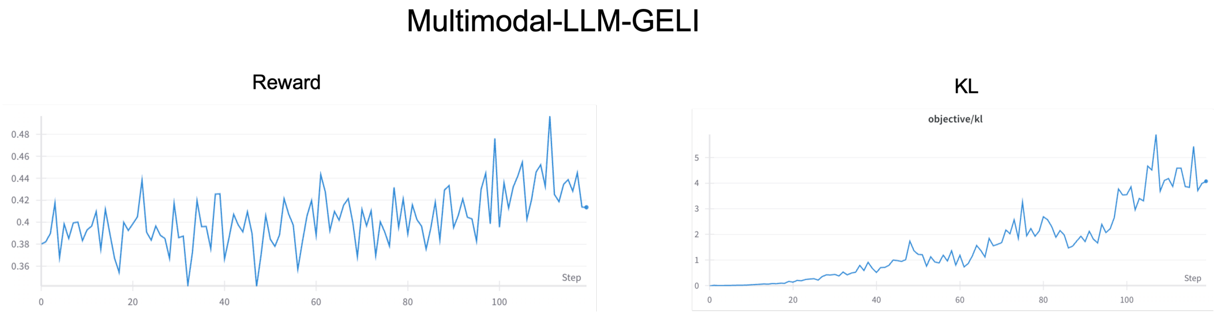


Figure 7: Multimodal-LLM-GELI RL Training during adaptation. Left: Reward scores over steps, Right: KL divergence over steps

J Consistency in LLM-based Reward Decomposition

To assess the consistency of the LLM-based reward decomposition, we queried an LLM for 10 conversations twice with temperature 1, and categorized each turn-level reward into three classes: positive (> 0), negative/neutral (0), or no annotation (NaN). Treating two independently sampled LLM reward decomposition outputs, we computed agreement across these 3-class labels. The agreement was 89.10% ± 5.37%, indicating that the model consistently identifies both the presence and polarity of rewards. This supports the reliability of LLM-generated signals as stable supervisory feedback.