RETROSPECTIVE LEARNING FROM INTERACTIONS

Zizhao Chen, Mustafa Omer Gul, Yiwei Chen, Gloria Geng, Anne Wu & Yoav Artzi
Department of Computer Science and Cornell Tech, Cornell University
{czz,momergul,annewu,yoav}@cs.cornell.edu {yc833,gcg46}@cornell.edu

ABSTRACT

Multi-turn interactions between large language models (LLMs) and users naturally include implicit feedback signals. If an LLM responds in an unexpected way to an instruction, the user is likely to signal it by rephrasing the request, expressing frustration, or pivoting to an alternative task. Such signals are task-independent and occupy a relatively constrained subspace of language, allowing the LLM to identify them even if it fails on the actual task. This creates an avenue for continually learning from interactions without additional annotations. We introduce RESPECT, a method to learn from such signals in past interactions via retrospection. We deploy RESPECT in a new multimodal interaction scenario, where humans instruct an LLM to solve an abstract reasoning task with a combinatorial solution space. Through thousands of interactions with humans, we show how RESPECT gradually improves task completion rate from 31% to 82%, all without any external annotation.

1 INTRODUCTION

Language models (LMs) often engage in multi-turn interactions with human users. Similar to human-human interactions, these interactions are naturally rich with implicit learning signals. If the LM fails to respond appropriately, the user is likely to follow with an expression of frustration, a rephrase of their intent, or maybe even completely pivot what they ask for. Similarly, if the LM does well, the user may express approval or simply continue to their next objective. Such responses can inform the LM of its performance, thereby creating an opportunity to learn through retrospection.

We study the efficacy of such signals, and how they can lead to a system that improves over time. We introduce RESPECT, a simple approach to learn from signals the model itself derives about its own past actions through retrospection of past interactions with human users. We deploy RESPECT in MULTIREF, a new multi-turn grounded interaction scenario, which requires models to display

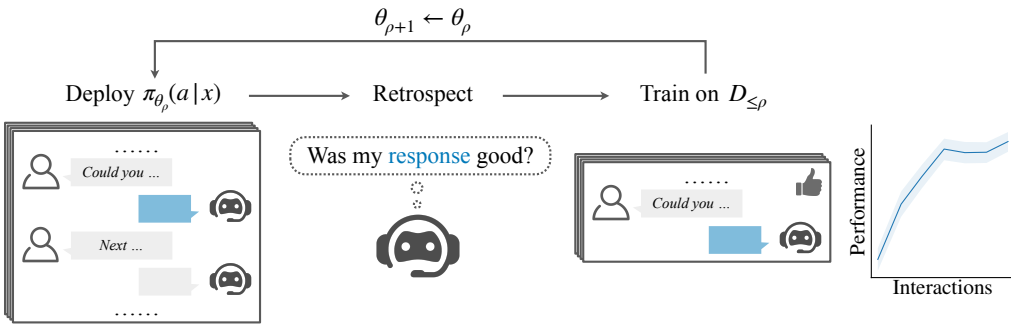

Figure 1: Learning via RESPECT. We deploy an LLM policy $\pi_{\theta_\rho}(a|x)$ in rounds ρ, to interact with users in multi-turn interactions. Following each round, the LLM reasons retrospectively about each of its actions (highlighted in blue) to decode feedback given the interaction context, including follow up utterances. After each round, the model is retrained using all data aggregated so far $D_{\leq\rho}$. The LLM improves over time without any external annotations. The plot on the right shows the performance curve in our experiments – the LLM improves from 31% to 82% task completion rate over six rounds.

complex abstract reasoning, and humans to gradually instruct models to accomplish sequences of goals to complete their tasks.

The key insight underlying RESPECT is that conversational implicit feedback signals occupy a relatively constrained subspace of natural language. Such signals can include direct approvals (e.g., *great!*) or signs of frustration (e.g., *not again*), and also more subtle cues, such as when the user rephrases their request. Critically, it is relatively simple to disentangle them from task performance. A human can easily figure out from such cues if they do well or not, even if they have little understanding about what they are asked for. It is this constrained nature that makes reasoning about such signals to be within the capacities of large language models (LLMs), even if they fail at the task at hand.

RESPECT utilizes this signal in a process where the model interacts with humans, and after interaction decodes feedback for each of its actions from the interaction context including the follow up utterances. Figure 1 illustrates this process. The model interacts with humans to accomplish tasks, retrospectively examines its own past interactions, and then re-trains. This process progresses in rounds, alternating between interaction and training, with the model improving over time. Critically, unlike common recipes for training from human feedback, RESPECT does not require any external annotation (Ouyang et al., 2022, RLHF) or even soliciting feedback from the users themselves (Suhr & Artzi, 2023).

We deploy RESPECT in MULTIREF over multiple rounds of grounded interactions with human use and re-training. We use IDEFICS2-8B (Laurençon et al., 2024) as our LLM, and experiment with multiple learning methods, including supervised learning, REINFORCE-style policy gradient (Williams, 1992; Kojima et al., 2021), and KTO (Ethayarajh et al., 2024). Across our experiments, we observe that IDEFICS2-8B effectively decodes feedback, even as it initially performs poorly in the same interactions. In our longest running experiment, we observe model task completion rate improves from 31% to 82%. Our code, data, and models are at `https://lil-lab.github.io/respect`.

2 TECHNICAL OVERVIEW AND NOTATION

We conduct continual learning studies by deploying our approach in MULTIREF, a new multi-turn grounded interaction scenario (Section 3). Overall, the study progresses in rounds, where the LLM policy is first deployed to interact with users and complete tasks, and the interactions are then used to re-train the policy. Our study involves multiple rounds, and our goal is to observe and evaluate the long-term dynamics of the process. This includes the robustness of our award decoding and training methods to the changing distribution of the data likely to be seen in an adaptive system in the wild. Section 3 describes our interaction scenario in detail, and Section 4 our learning method. First, we outline our problem of interest and its notation in abstract terms.

Task Notation The policy's task is to respond effectively to human utterances given in conversational context. Formally, let $\pi(a_t|x_t)$ be the policy that controls the listener behavior, with a_t an action string that represents the model response and x_t being the context on which the policy is conditioned, both at time t in the interaction. The context includes the instruction history up to and excluding time t, including current (i.e., at time $t-1$) and past speaker utterances, as well as any other relevant context in which the interaction takes place. As our learning progresses in rounds, we denote θ_ρ as the model parameters in round ρ, and π_{θ_ρ} as the parameterized policy.

Learning and Deployment We study a continual learning setup, where the learning signal is acquired from interactions of the deployed model with human speakers. Our study progresses in rounds (Figure 1). Each round ρ includes a deployment, followed by training. During deployment at round ρ, the model π_{θ_ρ} interacts with users. For each model action $\hat{a}_t \sim \pi_{\theta_\rho}(a|x_t)$, we record a tuple $(x_t, \hat{a}_t, p_t, \bar{f}_t)$, where x_t is the context given to the model at time t to predict action $\hat{a}_t$, p_t is the probability of $\hat{a}_t$ at the time of prediction, and $\bar{f}_t$ is the remainder of the interaction following $\hat{a}_t$. Critically, these interaction tuples contain no explicit feedback. We compute the implicit feedback $\hat{\gamma}_t$ using a feedback decoder $\phi(x_t, \hat{a}_t, \bar{f}_t)$, to obtain tuples $(x_t, \hat{a}_t, \hat{\gamma}_t, p_t)$. We experiment with multiple learning objectives using this feedback: supervised learning (SFT), policy gradient, and KTO.

Evaluation We measure the quality of the listener model $\pi_{\theta_\rho}(a_t|x_t)$ at each round ρ primarily by interaction success rates from live human-bot deployments. The same interactions are used to train the model for the next round. We track various characteristics of model behavior, such as number