NLP Reading Group
Feb 4, 2026

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Enhancing Personalized Multi-Turn Dialogue with Curiosity Reward

NeurIPS 2025

Yanming Wan[2], Jiaxing Wu[1], Marwa Abdulhai[4], Lior Shani[3], Natasha Jaques[1, 2]

[1]Google DeepMind
[2]University of Washington
[3]Google Research
[4]University of California, Berkeley

Presented by Sriram Sai Ganesh

Sections

1. Introduction
2. Related Work
3. CURIO
4. Experiments
5. Results

Introduction

Intuition:

Task: ***Personalized conversation*** –
User-specific conversational recommendations.

Predominant approaches:

- **RLHF:** treats all users as a homogeneous group.
- **Reward-modeling:** requires additional user-specific training.

Limitation: **<u>Existing approaches are limited in scope & require rich data in advance.</u>**

CURIO allows for online personalization:

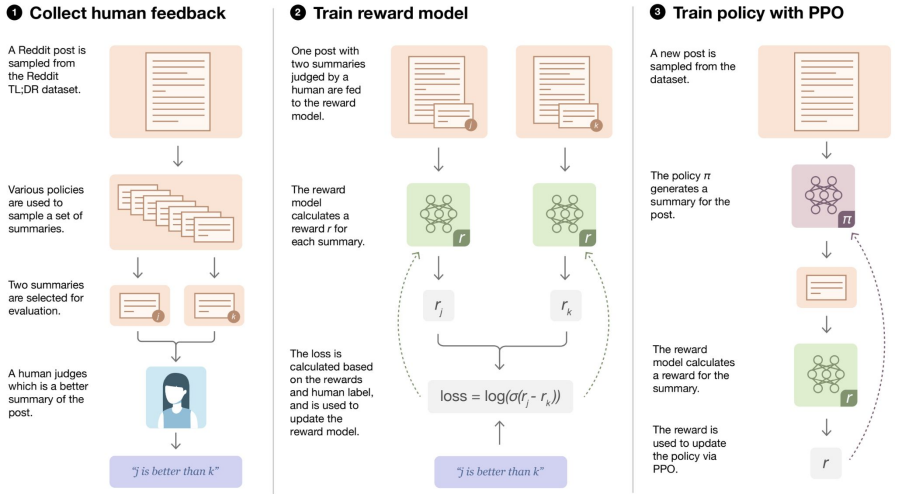
- LLM learns during the conversation.
- LLM adapts its interaction style for personalized conversations.

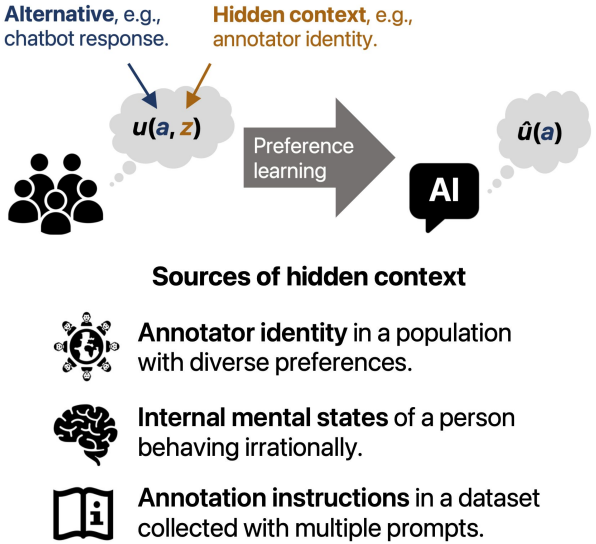
This paper:
Can we balance:

- exploration (learning about the user)
- exploitation (conversation rewards)?

Background

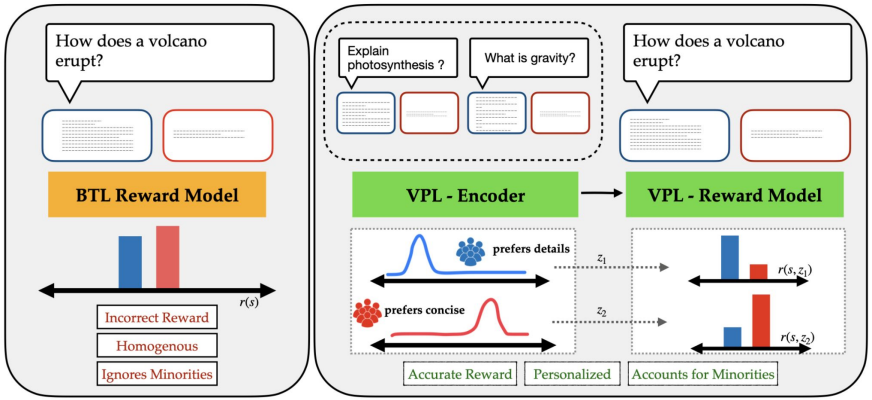
Reinforcement Learning from Human Feedback (RLHF)

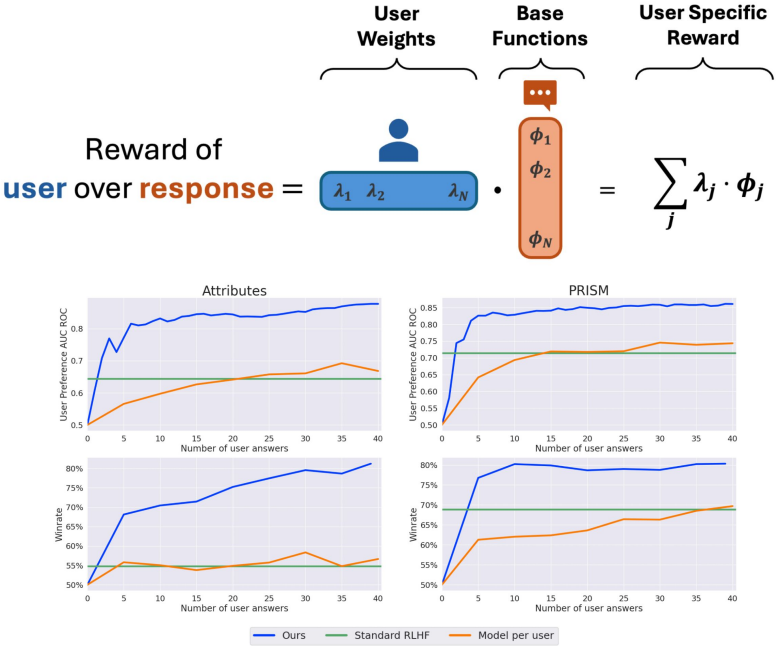




Background

Personalizing Conversation





CURIO

Overview

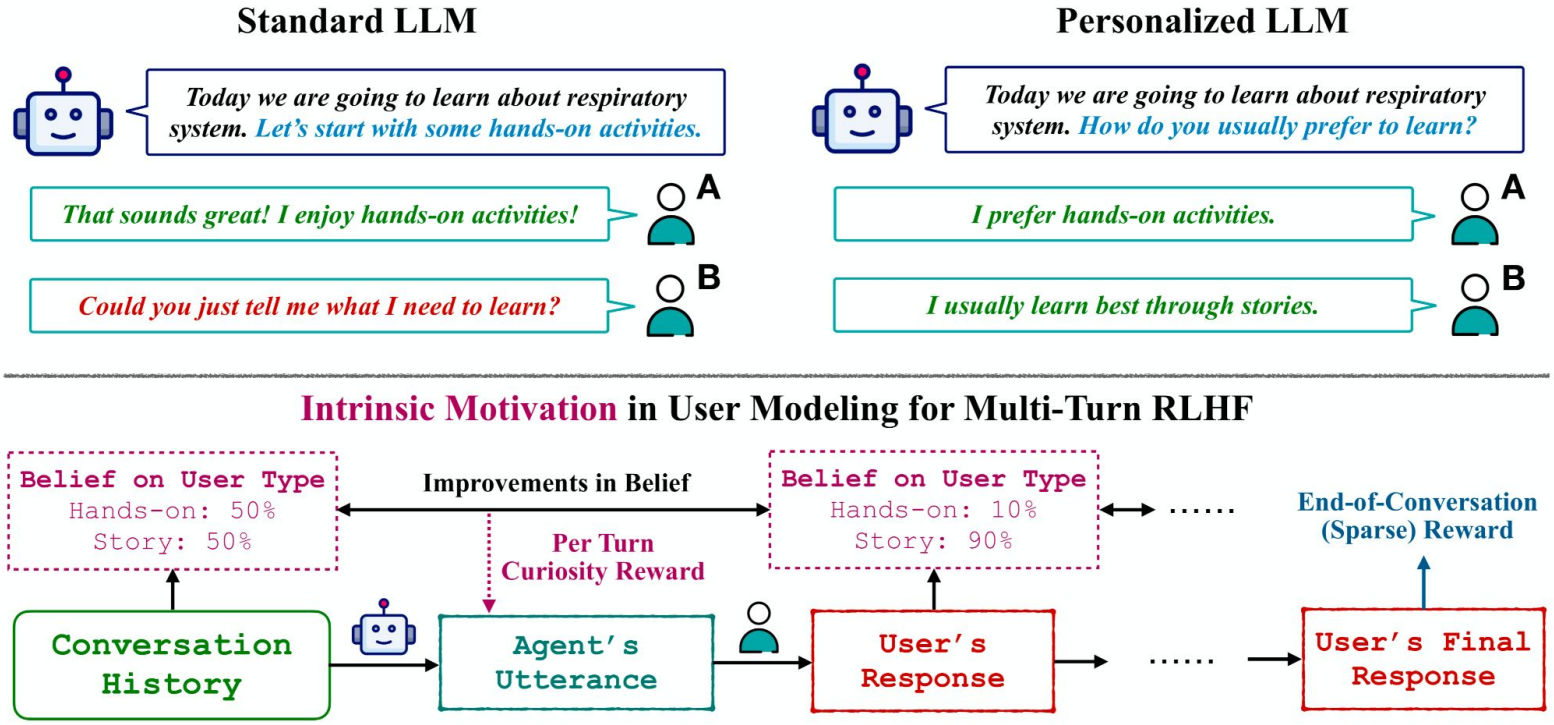


Figure 1

CURIO

User-conditioned RLHF

- **C**uriosity-driven **U**ser-modeling **R**eward as **I**ntrinsic **O**bjective
- Formulated as a **P**artially **O**bservable **M**arkov **D**ecision **P**rocess (POMDP)
- Defined by $(\tilde{\mathcal{S}}, \mathcal{U}, \mathcal{A}, \tilde{\mathcal{T}}, \tilde{\mathcal{R}}, \gamma)$

$u \in \mathcal{U} \rightarrow$ User type (hidden)

$\tilde{s}_t = \langle s_t, u \rangle \rightarrow$ current conversation rollout (LM $\cup$ user) + user type *u*

$a_t \in \mathcal{A} \rightarrow$ Action = response generated by LM.

$\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \rightarrow$ Reward function.

$\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}) \rightarrow$ Transition dynamics $\tilde{\mathcal{T}}(\tilde{s}_{t+1} \mid \tilde{s}_t, a_t) = \mathcal{T}(s_{t+1} \mid s_t, a_t, u)$

('~' = conditioned on user type *u*)

RLHF agent aims to maximize value $V^{\pi}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \mid \pi\right]$

CURIO

User-conditioned RLHF

- **C**uriosity-driven **U**ser-modeling **R**eward as **I**ntrinsic **O**bjective
- LLM in a POMDP environment:
  - Maintains a “belief function” $b_t \in \Delta(\mathcal{U})$
  - Updates belief at *t+1* as $b_{t+1}(u) \propto \mathcal{T}(s_{t+1} \mid s_t, a_t, u) b_t(u)$
  - Compute expected rewards over this (uncertain) belief distribution:

$$\mathcal{R}^b(s_t, b_t, a_t) = \sum_u b_t(u) \mathcal{R}(s_t, a_t \mid u)$$

  - Maximize cumulative reward:

$$V^\pi(s_0, b_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}^b(s_t, b_t, a_t) \mid \pi, s_0, b_0\right]$$

CURIO

Intrinsic Rewards

$$V^{\pi}(s_0, b_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}^b(s_t, b_t, a_t) \mid \pi, s_0, b_0\right]$$

- **Challenges**
  - Personalization success can typically only be evaluated **at the end of a conversation**. Sparse rewards => difficulty learning good early-turn actions.
  - Data imbalance among user groups => perform well only on majority group, **discouraging exploration** with other users.
- **Proposed Solution:**
  - Intrinsic Motivation (IM) to learn preferences out of “curiosity”, and then adapt.
  - Initialize a **user model** that takes *s_{t+1}* & predicts belief $b_{t+1}(u) := b(u|s_{t+1})$

Section 3.2 (Eqn. 3)

CURIO

Intrinsic Rewards

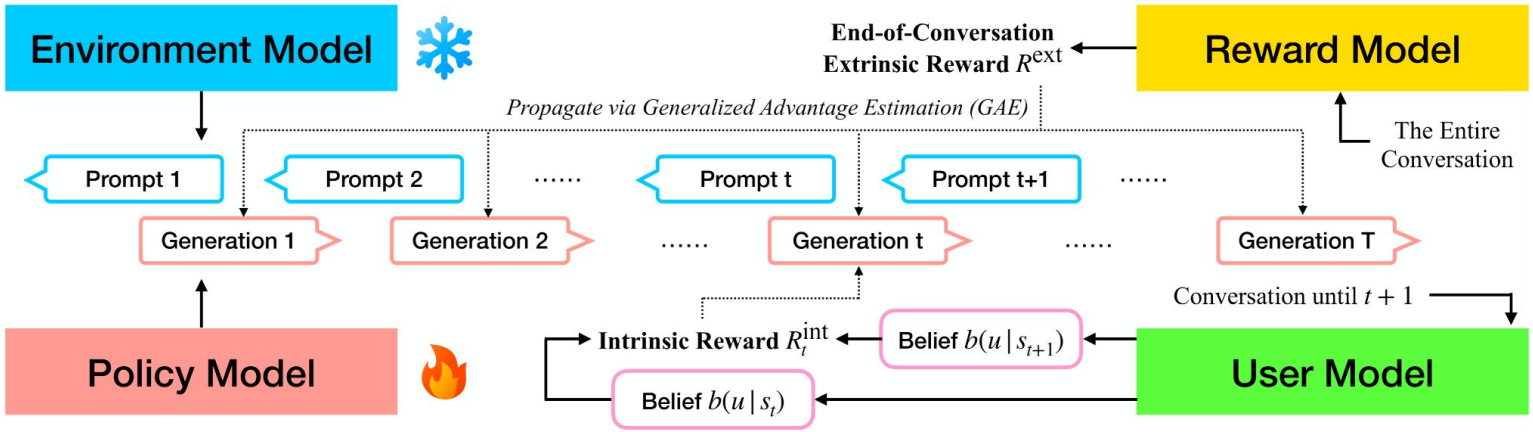
- **Proposed Solution:**
  - Intrinsic Motivation (IM) to learn preferences out of “curiosity”, and then adapt.
  - Initialize a **user model** that takes *s_{t+1}* & predicts belief $b_{t+1}(u) := b(u|s_{t+1})$
- Allows for turn-based intrinsic rewards:

$b_{t+1}(u^*) - b_t(u^*)$	• Improvement in accuracy over GT
$H(b_t) - H(b_{t+1})$	• Reduction in entropy

- CURIO framework:
  - 4 different LMs: *Policy, Environment, User, Reward.*

CURIO

Framework



Policy model: Engages in multi-turn conversation with *Environment model.*

Environment model: Simulates the human user.

User model: Predicts user type at each turn, based on *s_t*. These give us **turn-based rewards**.

Reward model: extrinsic end-of-conversation reward.

CURIO

Metrics

Designing intrinsic rewards:

1. **Potential-Based Reward Shaping (PBRS):** Widely used in RL, extended to POMDPs.
   - If ϕ defines a function over b_t $\phi : \Delta(\mathcal{U}) \to \mathbb{R}$
   - Reward by adding the difference between agent's belief at two timesteps:

$$r^b(s_t, b_t, a_t) = \mathcal{R}^b(s_t, b_t, a_t) + \gamma\phi(b_{t+1}) - \phi(b_t)$$

   - Does not affect optimal policy; optimizing r^b yields the same policy as V^π.
   - 3 candidates for ϕ to incentivize improvement in user prediction:

$$\phi_{\text{acc}}(b) = b(u^*), \quad \phi_{\text{log-acc}}(b) = \log b(u^*), \quad \phi_{\text{neg-ent}}(b) = -H(b) = \sum_u b(u) \log b(u),$$

   - Accuracy, log-accuracy and negative entropy of distribution of beliefs about the user.

CURIO

Metrics

Designing intrinsic rewards:

2. **Other reward shaping:**
   - Possible intrinsic curiosity rewards.
   - Do not guarantee optimality of policy learning.

Potential-based Reward Shaping		Other Reward Shaping	
DiffAcc	$\gamma b_{t+1}(u^*) - b_t(u^*)$	**Acc**	$b_{t+1}(u^*) - 1/\lvert\mathcal{U}\rvert$
DiffLogAcc	$\gamma \log b_{t+1}(u^*) - \log b_t(u^*)$	**Ent**	$\log \lvert\mathcal{U}\rvert - H(b_{t+1})$
DiffEnt	$H(b_t) - \gamma H(b_{t+1})$	**InfoGain**	$D_{\mathrm{KL}}[b_{t+1}(u) \lvert\rvert b_t(u)]$

CURIO

Research Questions

Personalization as the main objective:

1. Can CURIO improve performance on personalization tasks?
    - Conversational personalization: **Exercise Recommendation**
    - LM recommends exercise strategy tailored to user's lifestyle, health conditions etc.
    - Synthetic dataset. Users have 20 attributes, LM picks 1 of 8 strategies.
    - Requires multiple turns eliciting information before choosing a strategy.

Personalization as a component of a larger task:

2. Can CURIO effectively personalize conversations when it is not the ultimate objective?
3. How does user learning affect conversation quality?
    - Education Dialog dataset – LM selects 1 of 2 learning styles.
    - Evaluated on Personalization & Conversation Quality.

Figure 2

CURIO

Results

Success Rates (%) of different models over Exercise Recommendation:

Baseline		Other Reward Shaping			Potential-based Reward Shaping		
SFT	**MTRLHF**	**InfoGain**	**Ent**	**Acc**	**DiffEnt**	**DiffLogAcc**	**DiffAcc**
54.0	68.5(+14.5)	63.0(+9.0)	82.0(+28.0)	84.0(+30.0)	84.0(+30.0)	86.0(+32.0)	**87.5(+33.5)**

Table 1

CURIO

Results

Education Dialogue: Side-by-side Auto Eval Results on Personalization:

	Baseline	Other Reward Shaping			Potential-based Reward Shaping		
	MTRLHF	**InfoGain**	**Ent**	**Acc**	**DiffEnt**	**DiffAcc**	**DiffLogAcc**
MTRLHF	-	93.04	55.70	7.91	51.90	42.72	24.05
InfoGain	6.96	-	42.41	0.00	29.11	9.18	0.63
Ent	50.00	57.59	-	39.56	43.35	49.05	44.62
Acc	**92.09**	100.00	60.44	-	70.57	85.13	64.87
DiffEnt	48.10	70.89	55.06	29.43	-	40.51	34.49
DiffAcc	**57.28**	90.82	50.95	14.87	59.49	-	34.81
DiffLogAcc	**75.95**	99.37	55.38	35.13	65.51	65.19	-

Table 2

CURIO

Results

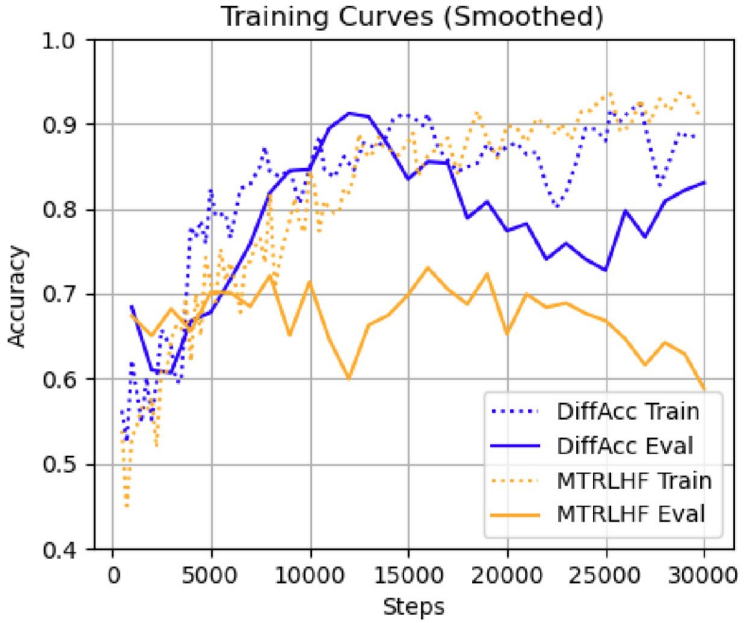
Education Dialogue: Side-by-side Auto Eval Results on Conversation Quality:

	Baseline	Other Reward Shaping			Potential-based Reward Shaping		
	MTRLHF	**InfoGain**	**Ent**	**Acc**	**DiffEnt**	**DiffAcc**	**DiffLogAcc**
MTRLHF	-	99.05	73.42	87.34	65.19	71.84	45.57
InfoGain	0.95	-	2.85	5.38	0.95	4.11	0.00
Ent	26.58	97.15	-	62.34	26.90	57.59	23.10
Acc	12.66	94.62	37.66	-	19.62	42.72	13.61
DiffEnt	34.81	99.05	73.10	80.06	-	73.73	31.65
DiffAcc	28.16	95.89	42.41	56.65	26.27	-	18.99
DiffLogAcc	**54.43**	**100.00**	**76.90**	**86.39**	**68.35**	**81.01**	-

Table 3

CURIO

Results



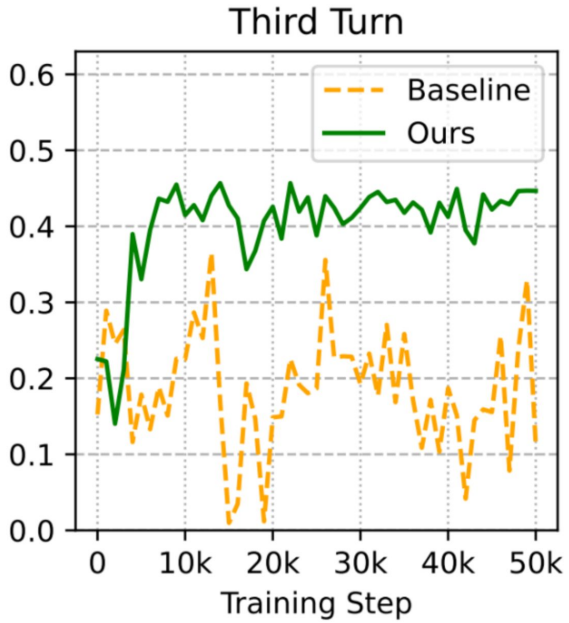


Figure 3, 4