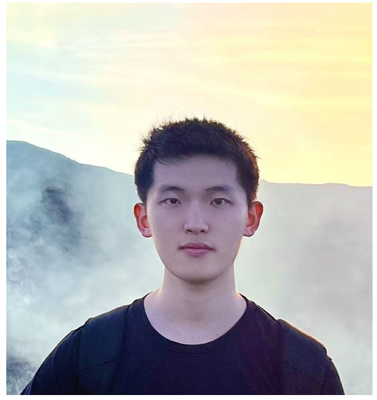
Enhancing Personalized Multi-Turn Dialogue with Curiosity Reward

Yanming Wan[2*†‡], **Jiaxing Wu**[1*†], **Marwa Abdulhai**[4], **Lior Shani**[3], **Natasha Jaques**[12]

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

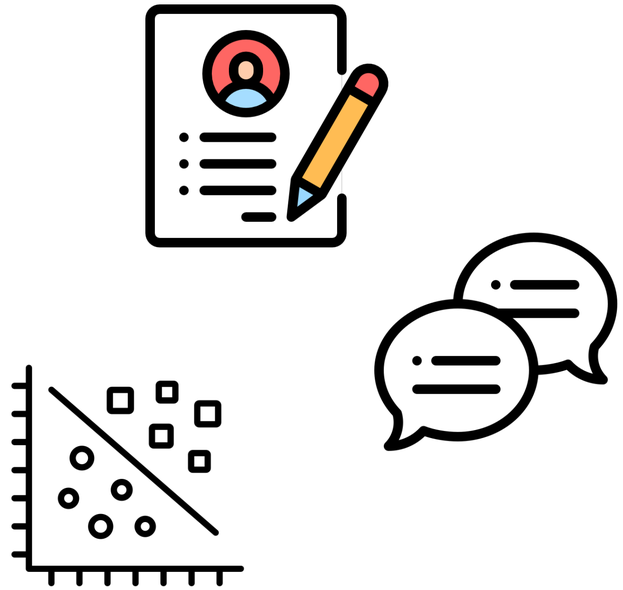
[*]Equal Contribution [‡]Work done during internship at Google DeepMind
[†]Correspondence to: <ymwan@cs.washington.edu, jxwu@google.com>



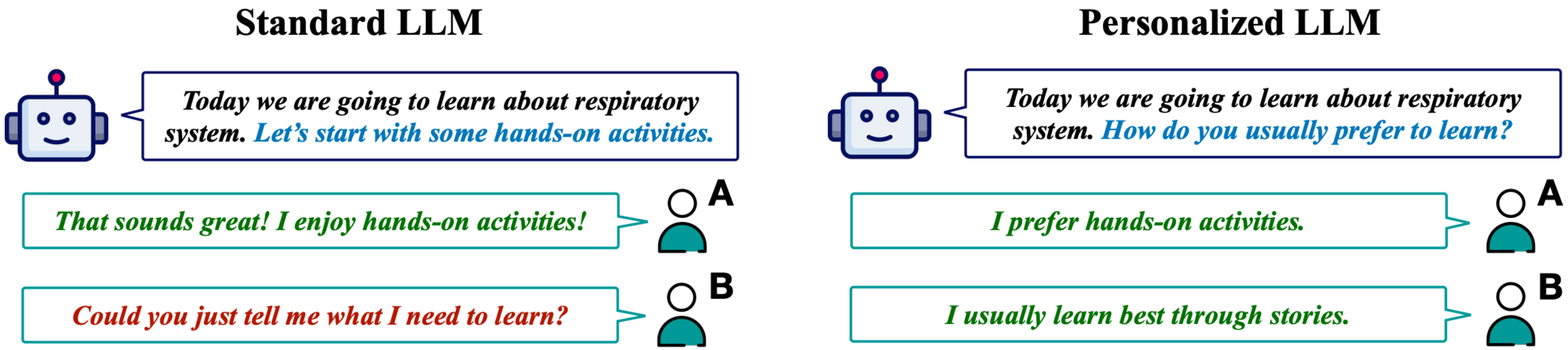


Personalization is missing in…

Gather user contexts beforehand?



Actively learn about user preferences!



Online Personalization

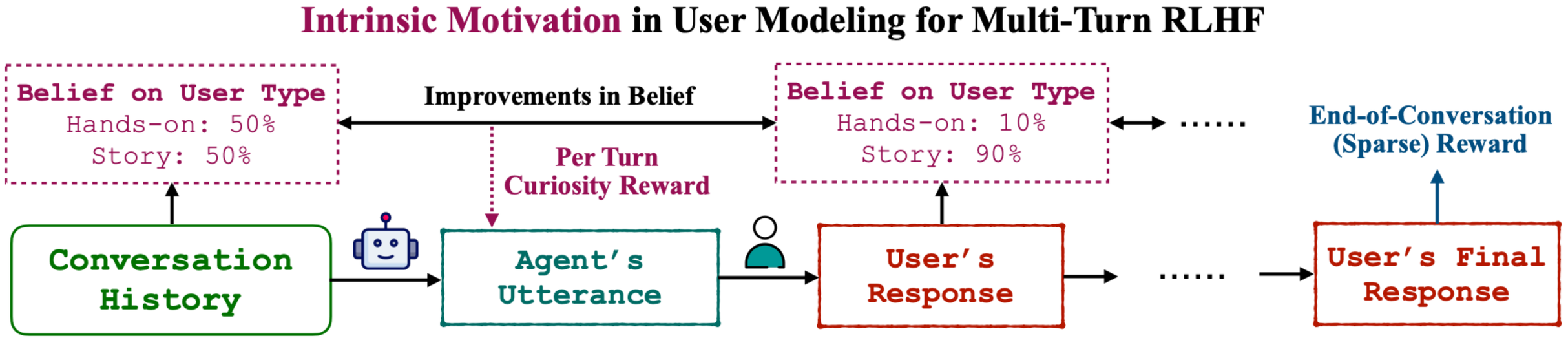
- **Key Idea:** train the LLM to conduct the conversation to get to know the user, enabling **online personalization**

Online Personalization

- **Key Idea:** train the LLM to conduct the conversation to get to know the user, enabling **online personalization**
- **Method:** Leverage a **user model**, and reward conversation turns that result in **improving the accuracy** of that user model.

Online Personalization

- **Key Idea:** train the LLM to conduct the conversation to get to know the user, enabling **online personalization**
- **Method:** Leverage a **user model**, and reward conversation turns that result in **improving the accuracy** of that user model.

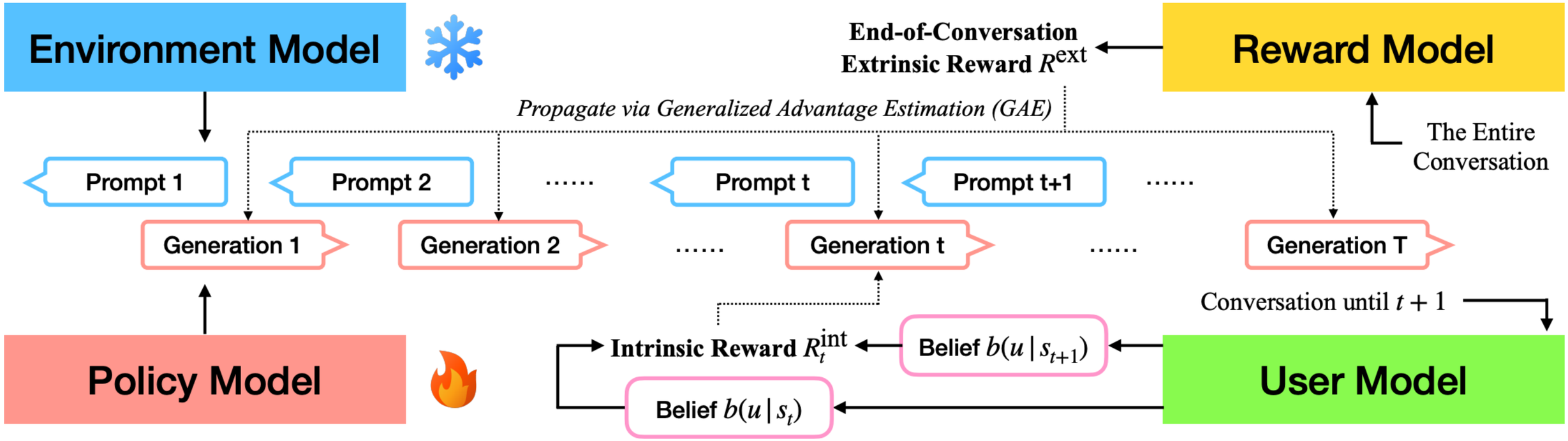


CURIO

Curiosity-driven User-modeling Reward as Intrinsic Objective

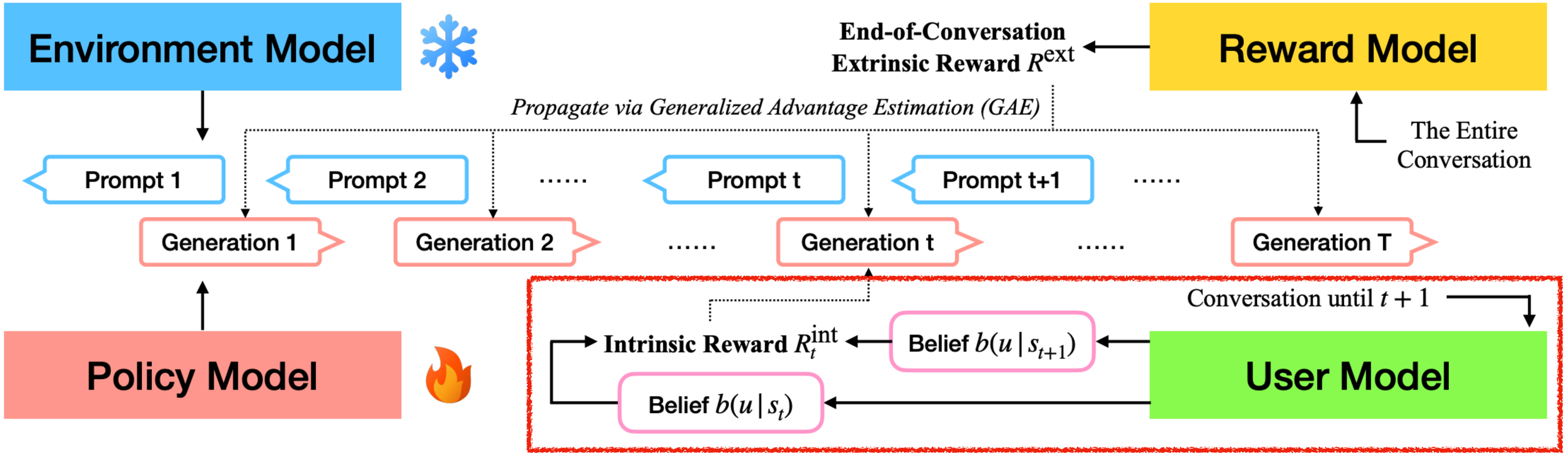
CURIO

Curiosity-driven User-modeling Reward as Intrinsic Objective



CURIO

Curiosity-driven User-modeling Reward as Intrinsic Objective



Theoretical Insight

Relationship with Potential-based Reward Shaping

Optimality guaranteed [1]

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

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

[1] Eck, A., Soh, L., Devlin, S. & Kudenko, D.. Potential-based reward shaping for pomdps. International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2013.

Theoretical Insight

Relationship with Potential-based Reward Shaping

Optimality guaranteed [1]

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

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

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**	$D_{\text{KL}}[b_{t+1}(u) \lvert\rvert b_t(u)]$

[1] Eck, A., Soh, L., Devlin, S. & Kudenko, D.. Potential-based reward shaping for pomdps. International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2013.

Results

Enhances personalization and reduces generalization gap

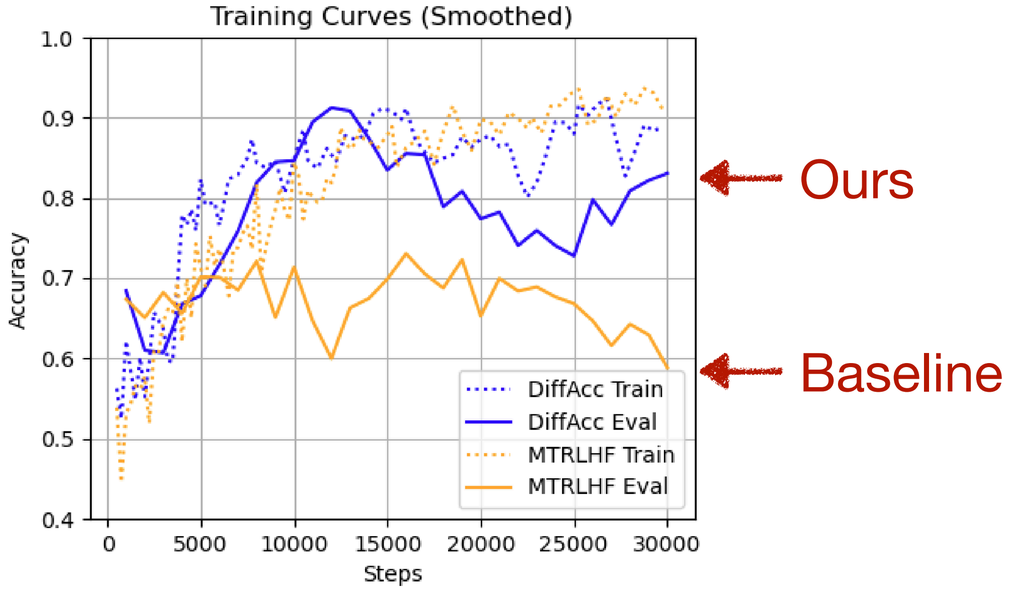
Baseline		Other Reward Shaping			Potential-based Reward Shaping		
SFT	**MTRLHF** [2]	**InfoGain**	**Ent**	**Acc**	**DiffEnt**	**DiffLogAcc**	**DiffAcc**
54.0	68.5(+14.5)	63.0(+9.0)	82.0(+28.0)	84.0(+30.0)	84.0(+30.0)	86.0(+32.0)	**87.5(+33.5)**

[2] Shani, L., Rosenberg, A., Cassel, A., Lang, O., Calandriello, D., Zipori, A., ... & Munos, R. (2024). Multi-turn reinforcement learning with preference human feedback. Advances in Neural Information Processing Systems, 37, 118953-118993.

Results

Enhances personalization and reduces generalization gap

Baseline		Other Reward Shaping			Potential-based Reward Shaping		
SFT	**MTRLHF** [2]	**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)**



[2] Shani, L., Rosenberg, A., Cassel, A., Lang, O., Calandriello, D., Zipori, A., ... & Munos, R. (2024). Multi-turn reinforcement learning with preference human feedback. Advances in Neural Information Processing Systems, 37, 118953-118993.

Results

Enhances personalization and reduces generalization gap

SFT

Agent: I can definitely help with that! To give you the best advice, I need to understand your lifestyle a bit better. Could you tell me your **name** and **age**? **[...]**

Agent: Strategy 6. **[Incorrect!]**

RLHF

Agent: What's your **name** and how old are you? **[...]**

Agent: Nice to meet you, Khloe! **What do you do** for now? **[...]**

Agent: That's a big time! Do you have any **hobbies or interests** you enjoy? **[...]**

Agent: Strategy 6. **[Incorrect!]**

CURIO

Agent: I can definitely help with that. To give you the best advice, could you tell me a bit more about yourself, like **your age** and **what you do for living**? **[...]**

Agent: Great. Do you have any **injuries or physical limitations** I should be be aware about? **[...]**

Agent: No problem at all. Do you prefer **indoors or outdoors**? **[...]**

Agent: Okay, how would you describe your **personality - introverted or extroverted**? **[...]**

Agent: Strategy 8. **[Correct!]**

Learning how to learn!

Results

Remains effective when personalization is relevant but not the goal

	Baseline	Other Reward Shaping			Potential-based Reward Shaping		
	MTRLHF	**InfoGain**	**Ent**	**Acc**	**DiffEnt**	**DiffAcc**	**DiffLogAcc**
MTRLHF [2]	-	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	-

[2] Shani, L., Rosenberg, A., Cassel, A., Lang, O., Calandriello, D., Zipori, A., ... & Munos, R. (2024). Multi-turn reinforcement learning with preference human feedback. Advances in Neural Information Processing Systems, 37, 118953-118993.

Conclusion

Curiosity-driven User-modeling Reward as Intrinsic Objective

- **Curiosity reward for Multi-turn RL** enables learning how to conduct conversations to actively get to know the users during the conversation
  - Enables online personalization and effective recommendations
  - ***Learns how to learn* about the user** rather than simply memorizing preferences of specific users