Inverse Reinforcement Learning Meets LLM Alignment

ACL 2025 Tutorial

Mihaela van der Schaar, Hao Sun

July. 2025, Vienna

van_der_Schaar
\ LAB

vanderschaar-lab.com


hs789@cam.ac.uk


@HolarisSun


sites.google.com/view/irl-llm

Content

- Part 1: Motivations
  - Breakthroughs on RL x LLMs
- Part 2: RL Meets LLMs: Forward and Inverse
- Part 3: Inverse: Learning Reward Models from Data
- Part 4: Forward: LLM Optimization with Reward Models
- Part 5: Insights from Sparse-Reward RL Literature
- Part 6: Infrastructure for RL x LLM Research

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 1: Motivations
- Part 2: RL Meets LLMs: Forward and Inverse
  - RL, MDP; Inverse RL, MDP\R
  - LLM alignment as Inverse RL
  - Why do we (always) need RMs?
- Part 3: Inverse: Learning Reward Models from Data
- Part 4: Forward: LLM Optimization with Reward Models
- Part 5: Insights from Sparse-Reward RL Literature
- Part 6: Infrastructure for RL x LLM Research

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 1: Motivations
- Part 2: RL Meets LLMs: Forward and Inverse
- Part 3: Inverse: Learning Reward Models from Data
  - Reward Modeling for RLHF
  - Reward Modeling for Math
- Part 4: Forward: LLM Optimization with Reward Models
- Part 5: Insights from Sparse-Reward RL Literature
- Part 6: Infrastructure for RL x LLM Research

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 1: Motivations
- Part 2: RL Meets LLMs: Forward and Inverse
- Part 3: Inverse: Learning Reward Models from Data
- Part 4: Forward: LLM Optimization with Reward Models
  - Optimization Algorithms
  - Challenges
- Part 5: Insights from Sparse-Reward RL Literature
- Part 6: Infrastructure for RL x LLM Research

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 1: Motivations
- Part 2: RL Meets LLMs: Forward and Inverse
- Part 3: Inverse: Learning Reward Models from Data
- Part 4: Forward: LLM Optimization with Reward Models
- Part 5: Insights from Sparse-Reward RL Literature
  - Hindsight Methods
  - Reward Shaping and Credit Assignment
  - Self-Play and Adversarial Learning
- Part 6: Infrastructure for RL x LLM Research

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 1: Motivations
- Part 2: RL Meets LLMs: Forward and Inverse
- Part 3: Inverse: Learning Reward Models from Data
- Part 4: Forward: LLM Optimization with Reward Models
- Part 5: Insights from Sparse-Reward RL Literature
- Part 6: Infrastructure for RL x LLM Research
  - AReaL: A Super Fast RL System for LLMs (by Prof. Yi Wu)

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Part 1:

Motivations

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Success of Large-Scale RL

- Alpha-Go/Zero/Star/Tensor/Chip/Dev
- OpenAI-Five

Reward ⬆️⬆️

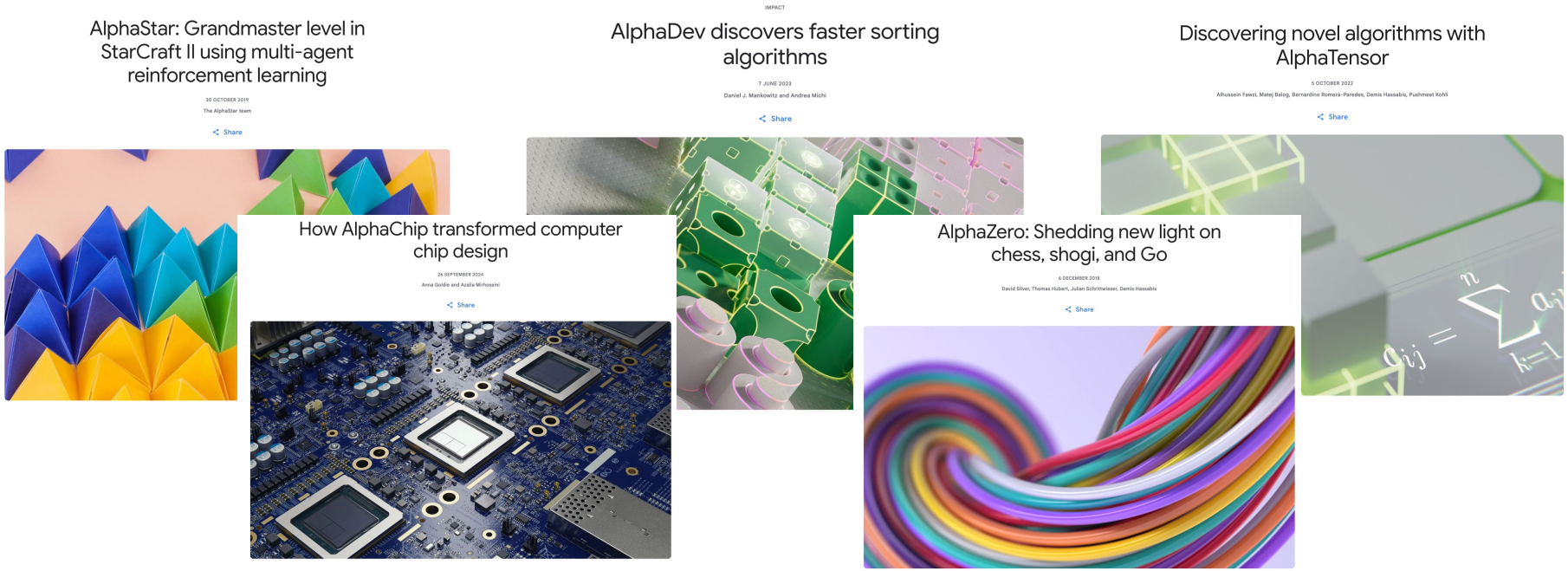


Image source: deepmind.google/discover

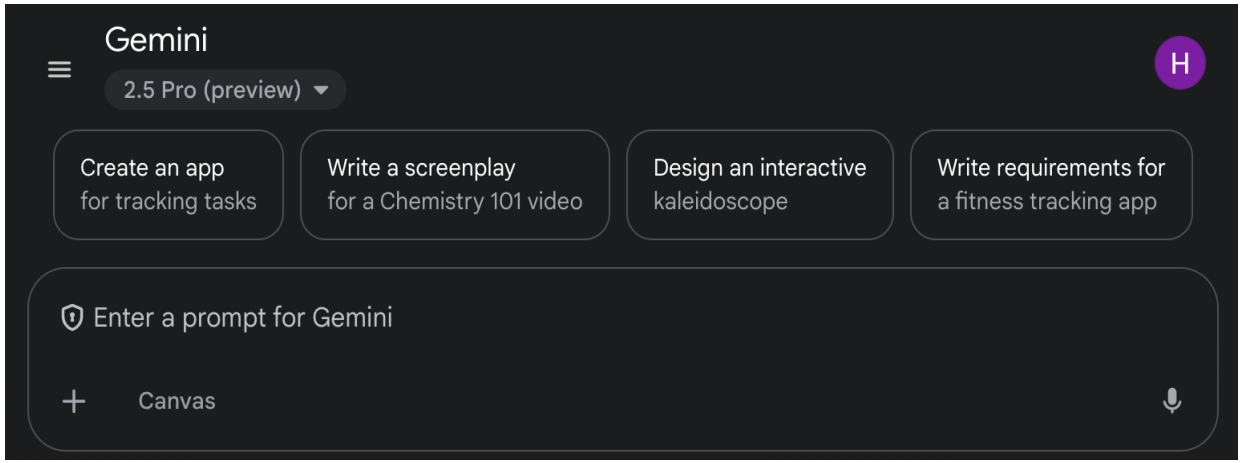
openai.com/index/openai-five/

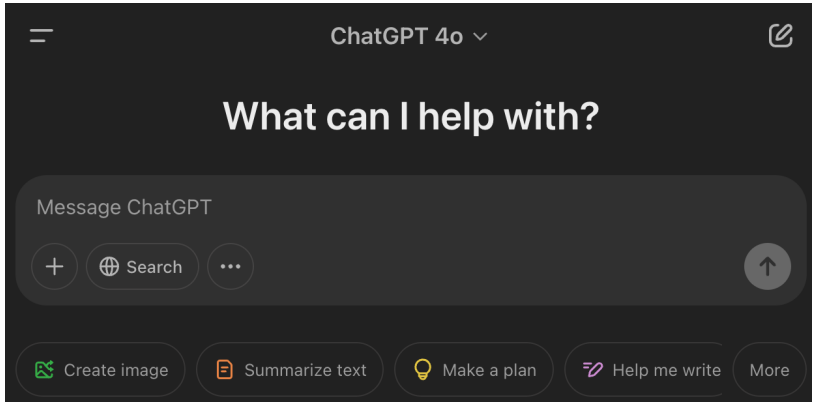
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Success of Data-Driven Large Models

- Image/Video Gen
- Large Language Models

Video source: openai.com/sora/





hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Success from Both Sides

Language (generative) Models

- Understanding and generating
- Fast adaptation to new tasks

RL

- Explore new knowledge
- Super-human (expert) performance
- Keep improving

[1] Talk: David Silver - Towards Superhuman Intelligence - RLC 2024

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Combining the Success?

Language Models	*RL*

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Combining the Success?

Language Models

RL

- RL agents can be the best Go/StarCraft/Dota2 player
- But learning from RL agents is non-trivial

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Combining the Success?

Language Models | ***RL***

- RL agents can be the best Go/StarCraft/Dota2 player
- But learning from RL agents is non-trivial

- ***LLM for RL***:
  Opportunity for more interpretable machine intelligence to assist, empower and inspire human.

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Motivation: Combining Success

Language Models *RL*

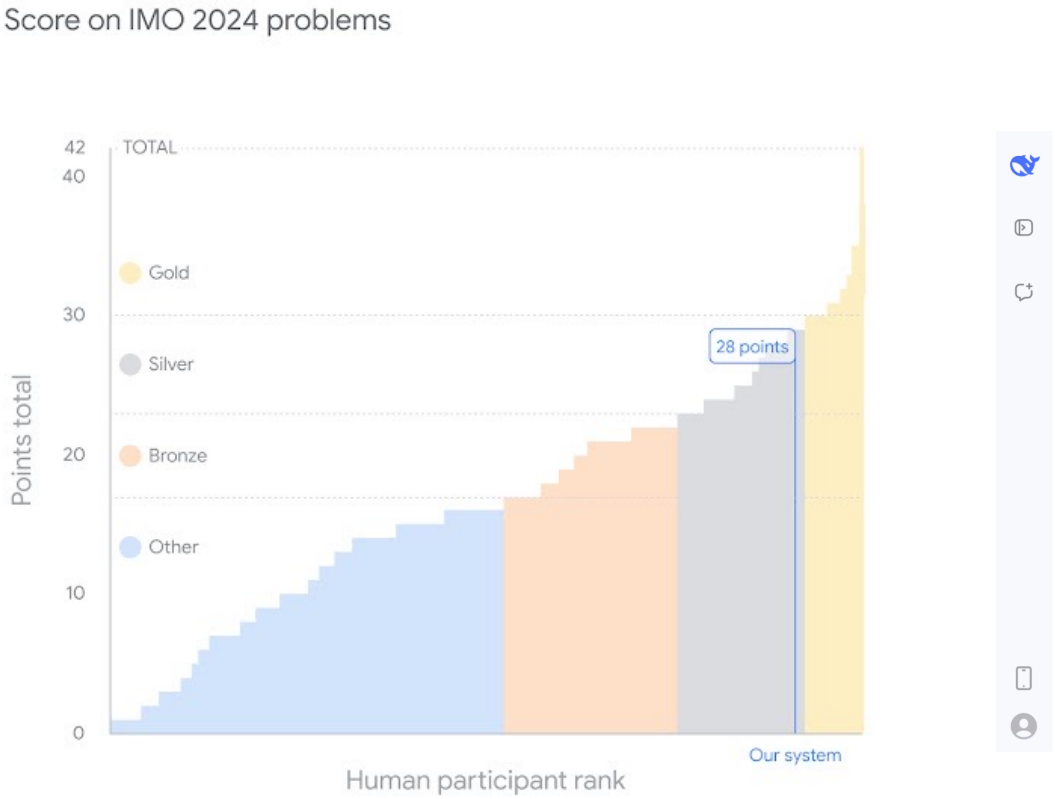
- RL to improve performance of LLMs
  - Explore new knowledge
  - Keep improving
  - Super-human performance

- ***LLM Alignment (incl. Agentic)****: to ensure its outputs are aligned with human intent, ethical principles, and task-specific requirements.*

hs789@cam.ac.uk

Success Case 1: Math/Reasoning

- AlphaProof & AlphaGeometry 2 LLM – o1/ DeepSeek-R1



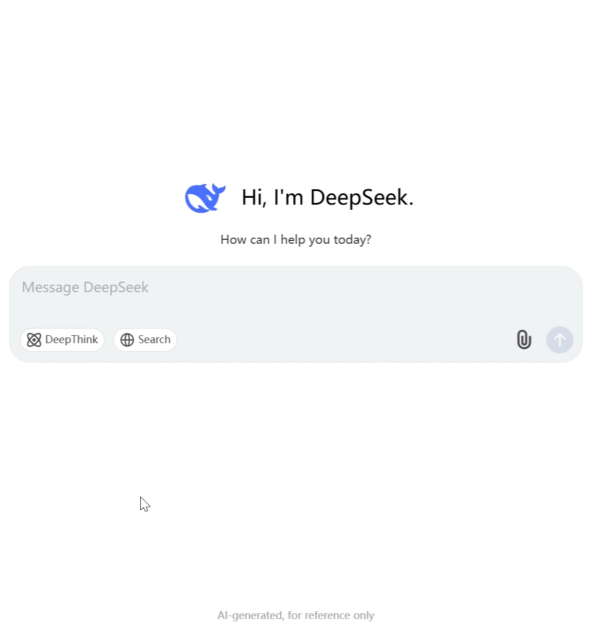


Image source: deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

hs789@cam.ac.uk

Success Case 1: Math/Reasoning

- AlphaProof & AlphaGeometry 2

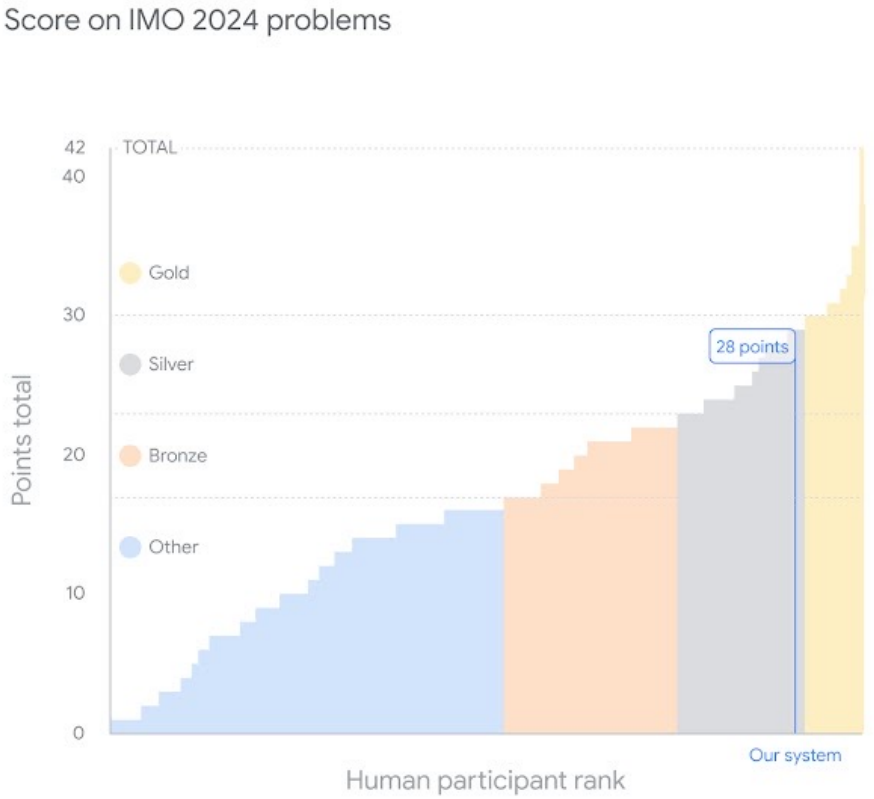
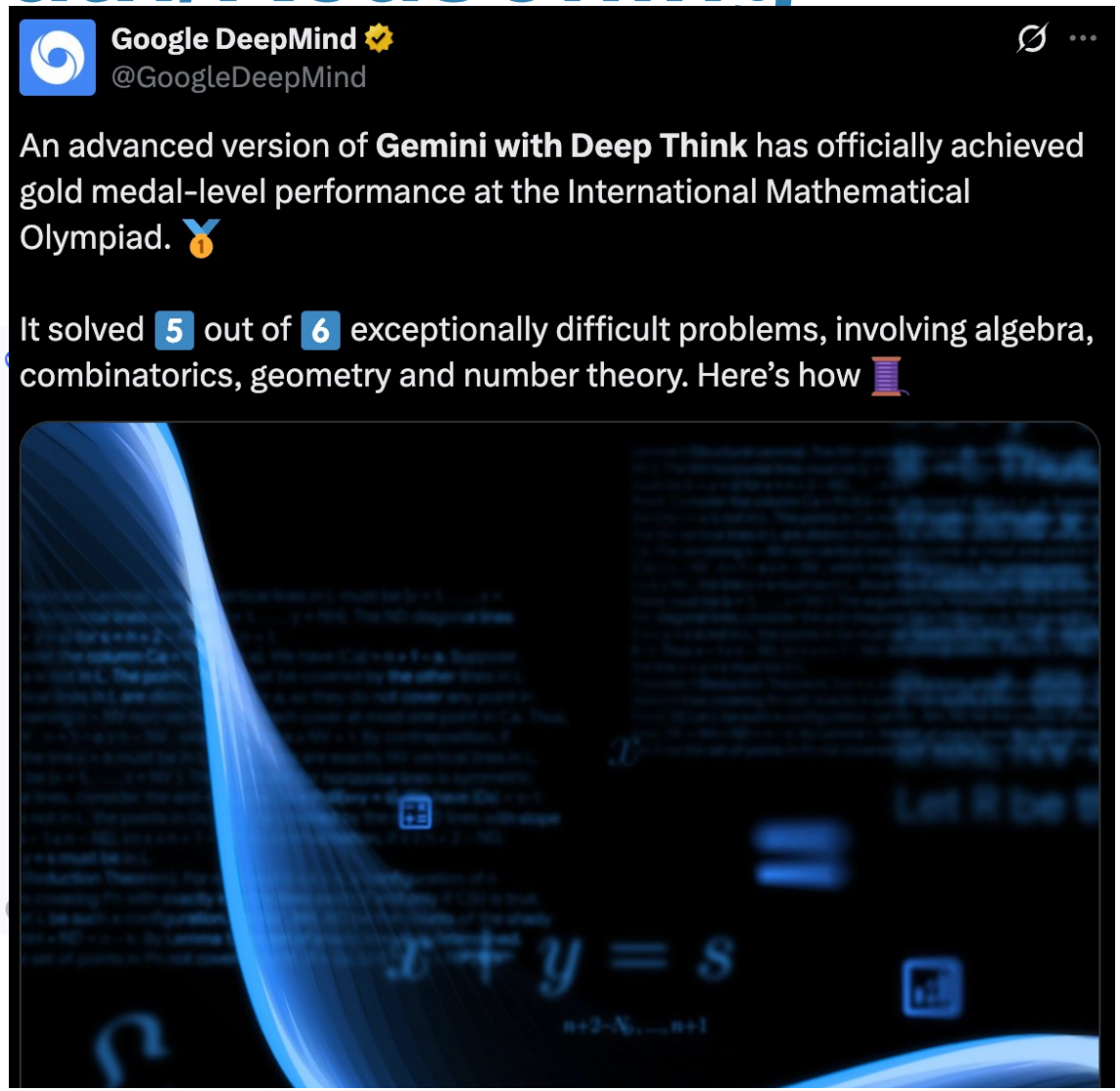


Image source: deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

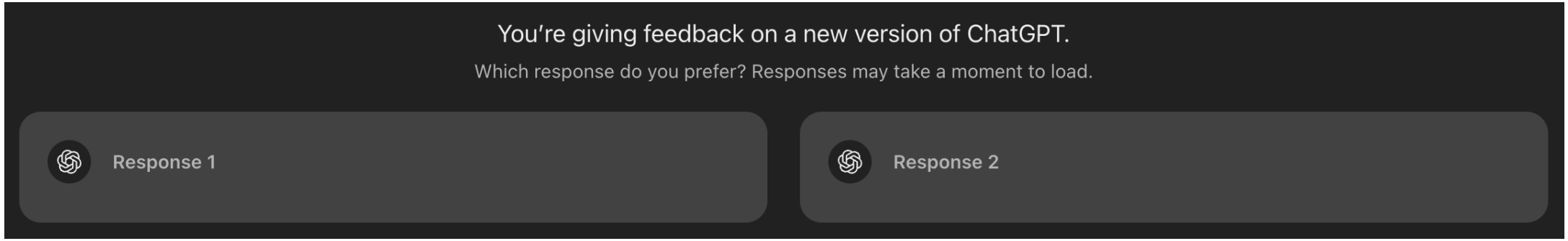


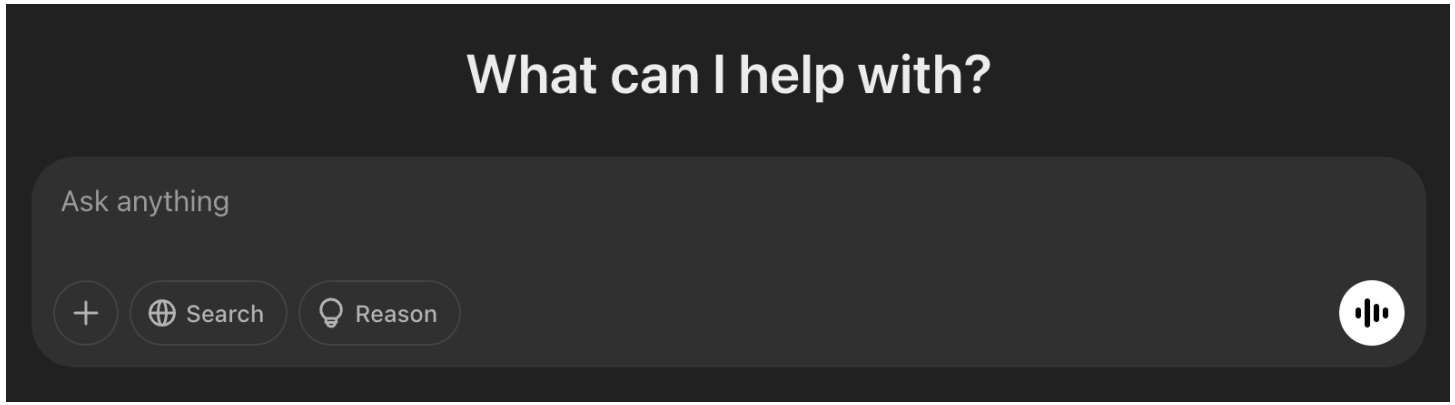
UNIVERSITY OF CAMBRIDGE

hs789@cam.ac.uk

Success Case 2: Chat

- ChatGPT (keep improving with user feedback)





hs789@cam.ac.uk

Motivating Questions

- Can we have a unified framework to understand the current success of RL in LLMs?
- How to extend the success of RL x LLM into more general tasks?
- What are the challenges and potential solutions?

hs789@cam.ac.uk

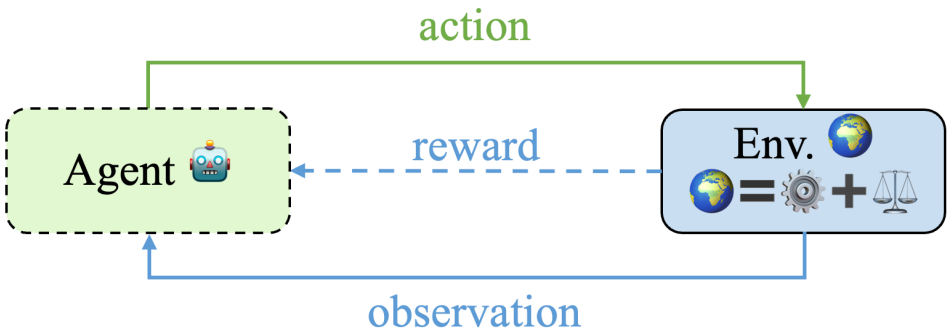
Part 2:

LLM Generation beyond Imitation: Forward and Inverse RL

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL: Learning through Interactions

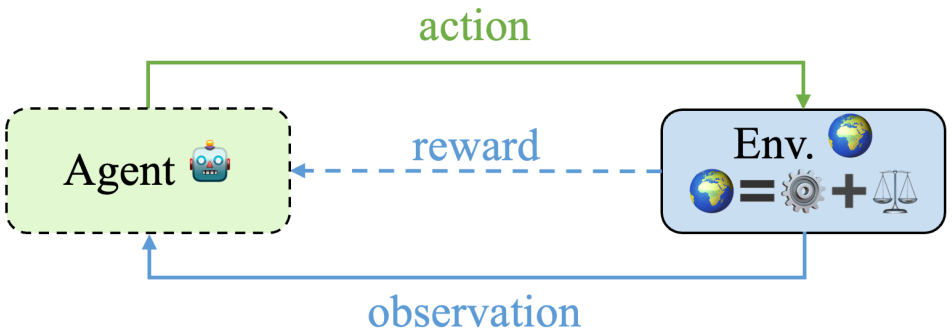
- Learning a policy to maximize the long-term return by interacting with the environment.



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL: Learning through Interactions

- Learning a policy to maximize the long-term return by interacting with the environment.
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL: Learning through Interactions

- Learning a policy to maximize the long-term return by interacting with the environment.
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$

$$\mathcal{J}(\pi(a|s))$$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL: Learning through Interactions

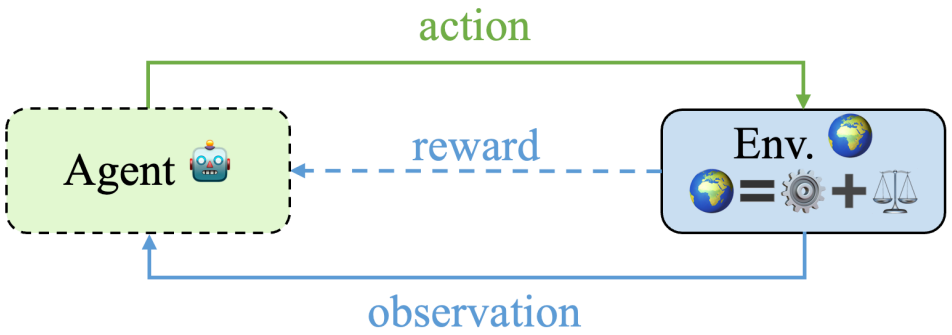
- Learning a policy to maximize the long-term return by interacting with the environment.
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$

$\mathcal{J}(\pi(a|s))$ $\sum_t \gamma^t r(s_t, a_t)$

RL: Learning through Interactions

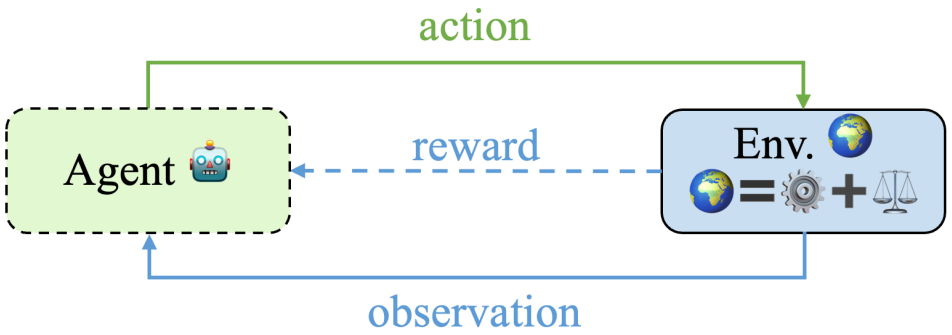
- Learning a policy to maximize the long-term return by interacting with the environment.
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$

$\mathcal{J}(\pi(a|s))$ $\mathbb{E}_{\rho_0, P}$ $\sum_t \gamma^t r(s_t, a_t)$



RL: Learning through Interactions

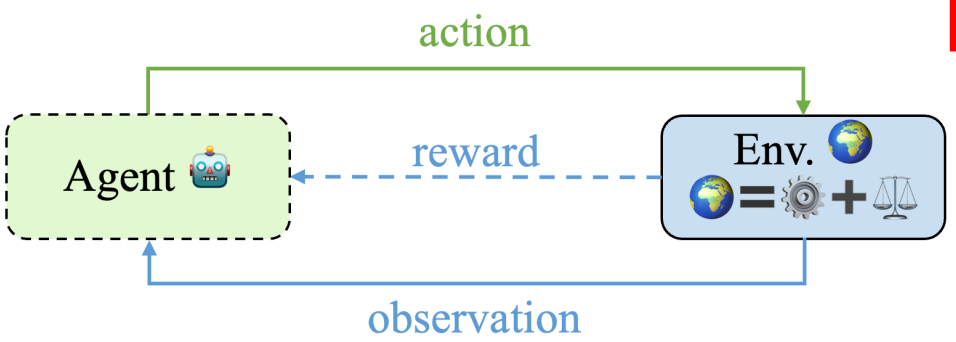
- Learning a policy to maximize the long-term return by interacting with the environment.
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- $\max_{\pi} \mathcal{J}(\pi(a|s)) = \max_{\pi} \mathbb{E}_{\rho_0, P, \pi}[\sum_t \gamma^t r(s_t, a_t)]$



RL: Learning through Interactions

- Learning a policy to maximize the long-term return by interacting with the environment.
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- $\max_{\pi} \mathcal{J}(\pi(a|s)) = \max_{\pi} \mathbb{E}_{\rho_0, P, \pi}[\sum_t \gamma^t r(s_t, a_t)]$

Formal Objective of RL



RL: Learning through Interactions

- How to optimize?

$$\max_{\pi} \mathcal{J}(\pi(a|s)) = \max_{\pi} \mathbb{E}_{\rho_0, P, \pi}[\sum_t \gamma^t r(s_t, a_t)]$$

- The core idea is “simple”:

discover and repeat successful trajectories/actions

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL: Learning through Interactions

- How to optimize?

$$\max_{\pi} \mathcal{J}(\pi(a|s)) = \max_{\pi} \mathbb{E}_{\rho_0, P, \pi}[\sum_t \gamma^t r(s_t, a_t)]$$

- The core idea is "simple":

discover and repeat successful trajectories/actions
explore exploit

RL: Learning through Interactions

- *Some* RL algorithms can be better than others in *some* tasks
- There is ***no silver bullet in RL***

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL: Learning through Interactions

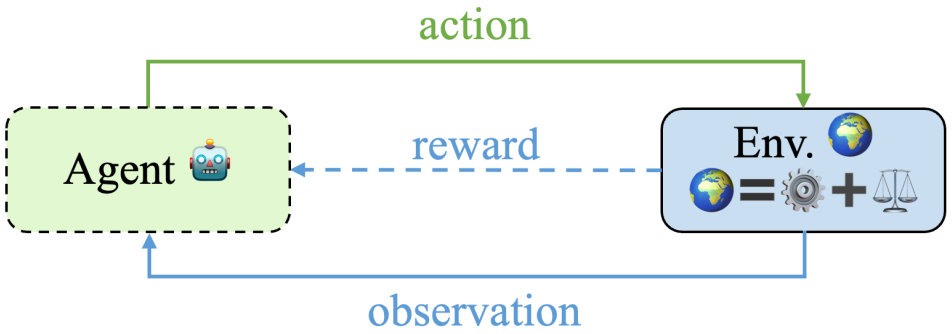
- *Some* RL algorithms can be better than others in *some* tasks
- There is ***no silver bullet in RL***
- e.g.,

Atari:	DQN
AlphaZero:	MCTS, Self-Play
OpenAI Five:	PPO, Self-Play
Robotics:	SAC, DPG
Multi-Goal:	Hindsight Methods
DeepSeek-r1:	GRPO
RLHF:	PPO, DPO, REINFORCE

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Generation as an MDP?

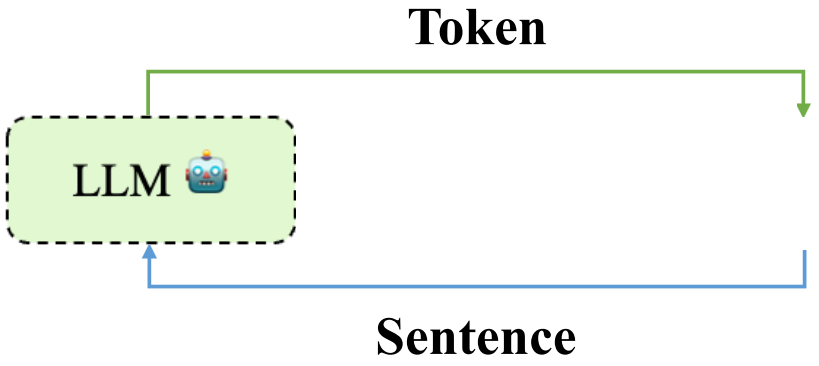
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

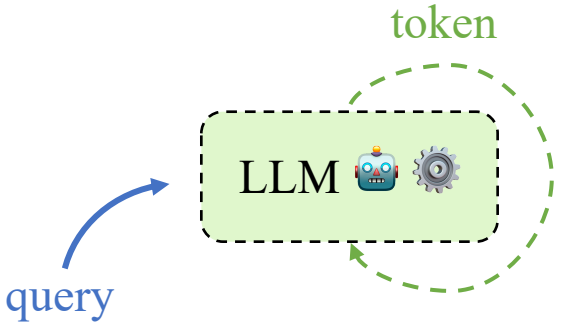
LLM Generation as an MDP?

- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- S: Current sentence
- A: Tokens (or their combinations)



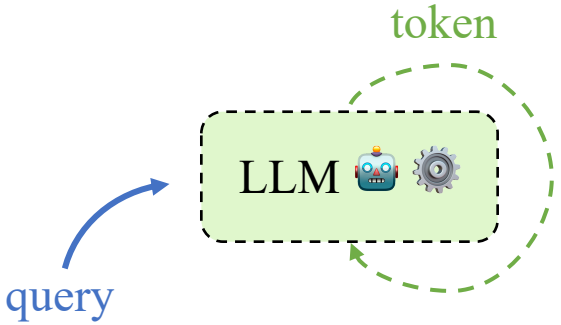
LLM Generation as an MDP?

- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- S: Current sentence
- A: Tokens (or their combinations)
- P: Concatenation of tokens
- ρ_0: Prompt/Query distribution



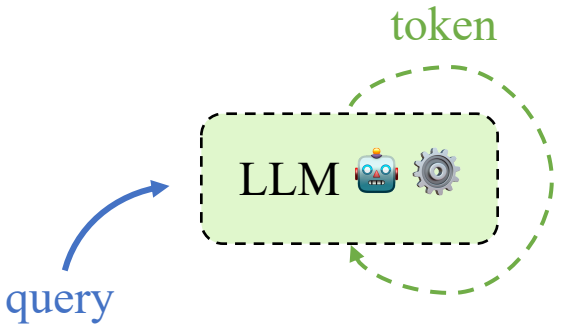
LLM Generation as an MDP\R

- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- S: Current sentence
- A: Tokens (or their combinations)
- P: Concatenation of tokens
- ρ_0: Prompt/Query distribution
- ? R: *(Data-Driven)*



LLM Generation as an MDP\R

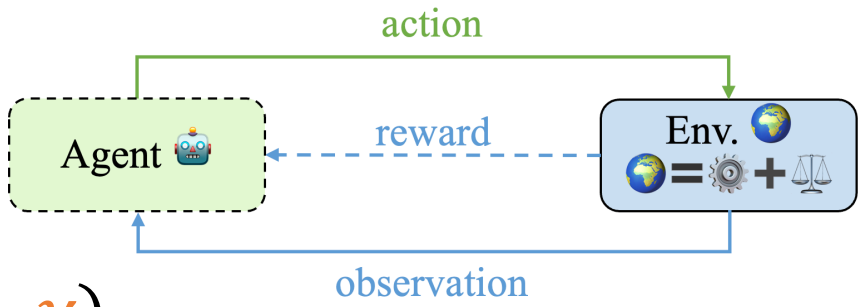
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- S: Current sentence
- A: Tokens (or their combinations)
- P: Concatenation of tokens
- ρ_0: Prompt/Query distribution
- ? R: *(Data-Driven)*
- γ: ≤ 1 (e.g., =1: correct is enough / <1: correct in short answer)



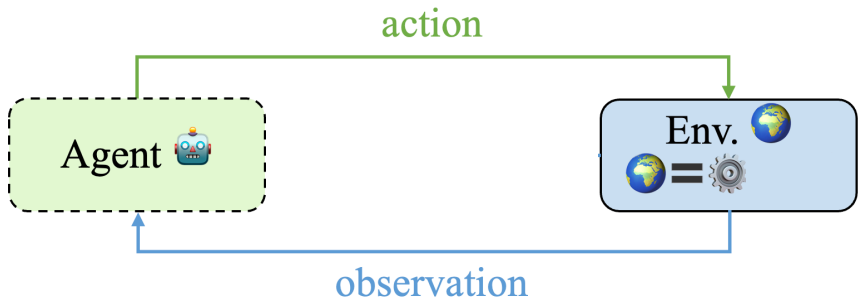
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

How to Learn in MDP\R

- In MDPs ($S, A, P, \rho_0, R, \gamma$), we maximize cumulative return



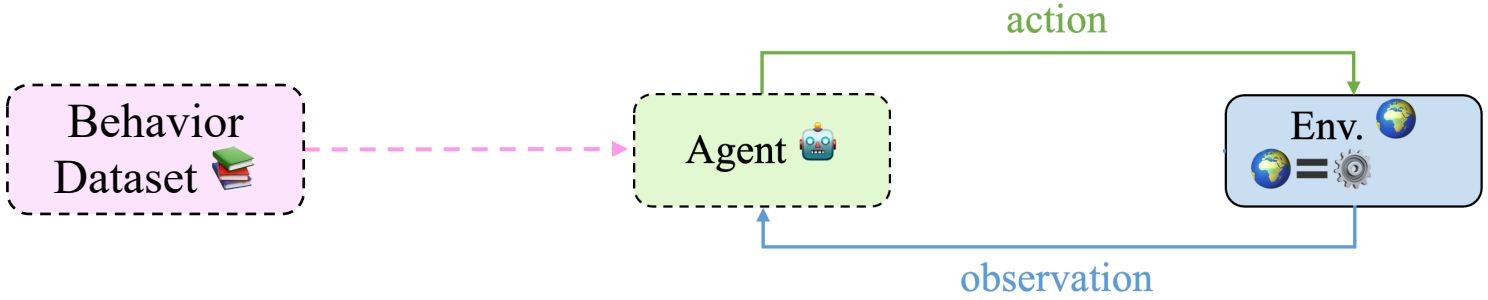
- In MDP\Rs (S, A, P, ρ_0, γ), how (what) to learn?



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Learning from Behavior Datasets

- MDP\R (S, A, P, ρ_0, γ), how (what) to learn?



- Learning from a *behavior dataset*

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Examples

- Learning from a *behavior dataset*
  - 1. Reward function is hard to define

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Examples

- Learning from a *behavior dataset*
  - 1. Reward function is hard to define
    - ALVINN [Pomerleau, 1988]

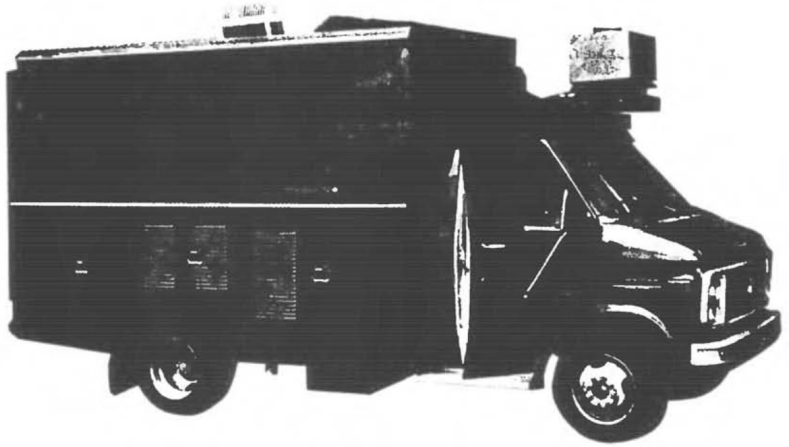


Figure 3: NAVLAB, the CMU autonomous navigation test vehicle.

Examples

- Learning from a *behavior dataset*
  - 1. Reward function is hard to define
    - ALVINN [Pomerleau, 1988]
    - Imitating behaviors [Hayes & Demiris, 1994]

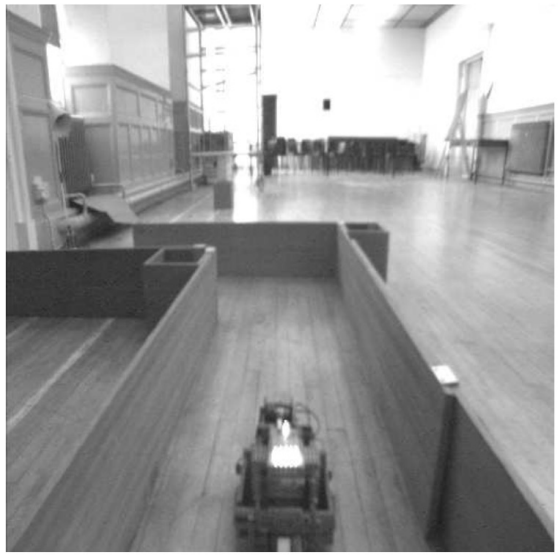


Fig. 3.: View of the teacher and part of the maze as seen by Ben Hope.

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Examples

- Learning from a *behavior dataset*
  - 1. Reward function is hard to define
    - ALVINN [Pomerleau, 1988]
    - Imitating behaviors [Hayes & Demiris, 1994]
    - Complex skills [Peng et al., 2016]

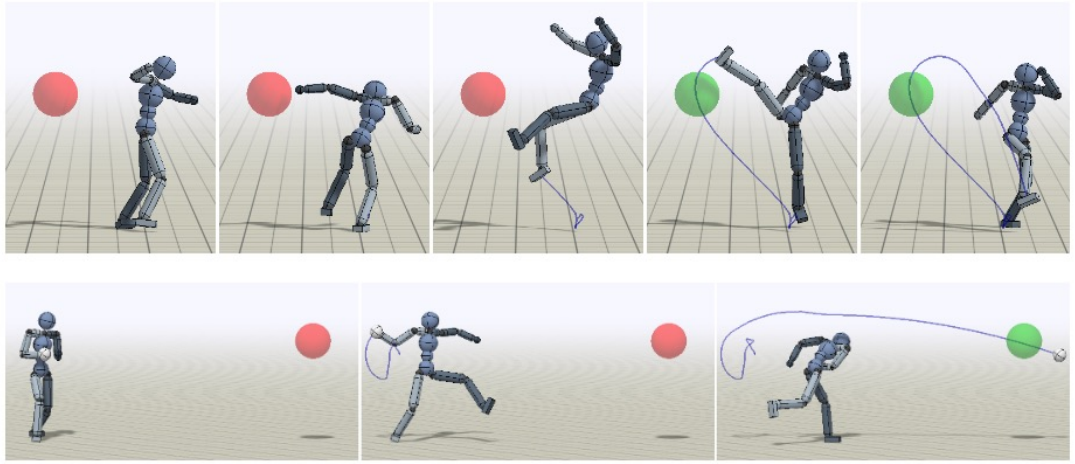


Fig. 7. **Top:** Spinkick policy trained to strike a target with the character's right foot. **Bottom:** Baseball pitch policy trained to throw a ball to a target.

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Examples

- Learning from a *behavior dataset*
  - 1. Reward function is hard to define
    - ALVINN [Pomerleau, 1988]
    - Imitating behaviors [Hayes & Demiris, 1994]
    - Complex skills [Peng et al., 2016]
    - RLHF: metrics are hard to quantify otherwise [Stiennon et al., 2020] [Bai et al., 2022]

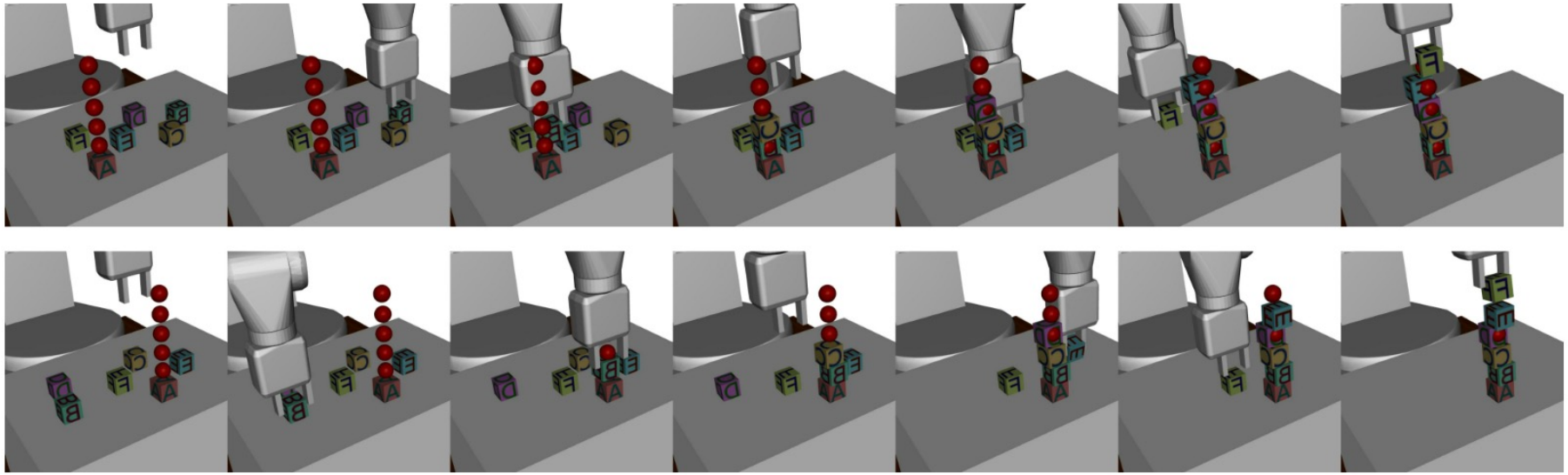
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Examples

- Learning from a *behavior dataset*
  - 1. Reward function is hard to define
    - ALVINN [Pomerleau, 1988]
    - Imitating behaviors [Hayes & Demiris, 1994]
    - Complex skills [Peng et al., 2016]
    - RLHF: metrics are hard to quantify otherwise [Stiennon et al., 2020] [Bai et al., 2022]
  - 2. Reward signal is too sparse (e.g., win a game of Go/StarCraft/Dota2)
    - AlphaGo/AlphaStar/OpenAI Five

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

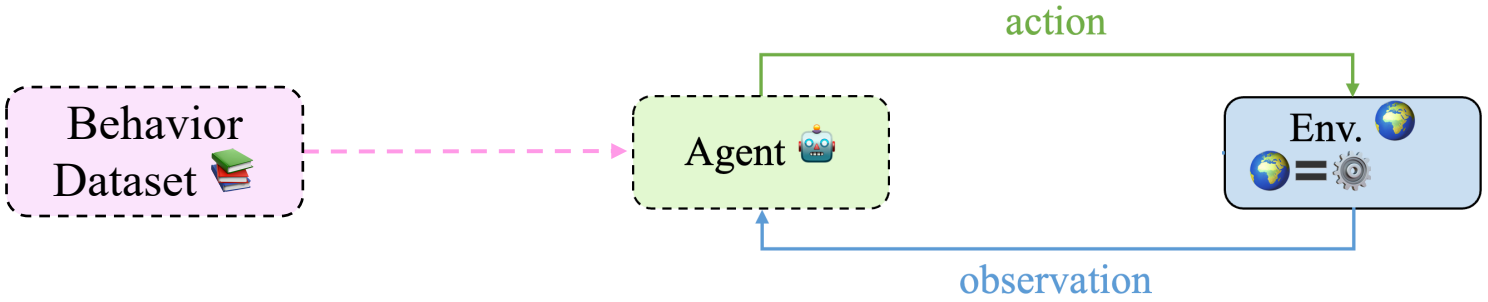
Examples



- 2. Reward signal is too sparse (e.g., win a game of Go/StarCraft/Dota2)
  - AlphaGo/AlphaStar/OpenAI Five
  - Robotics Control [Nair et al., 2017]

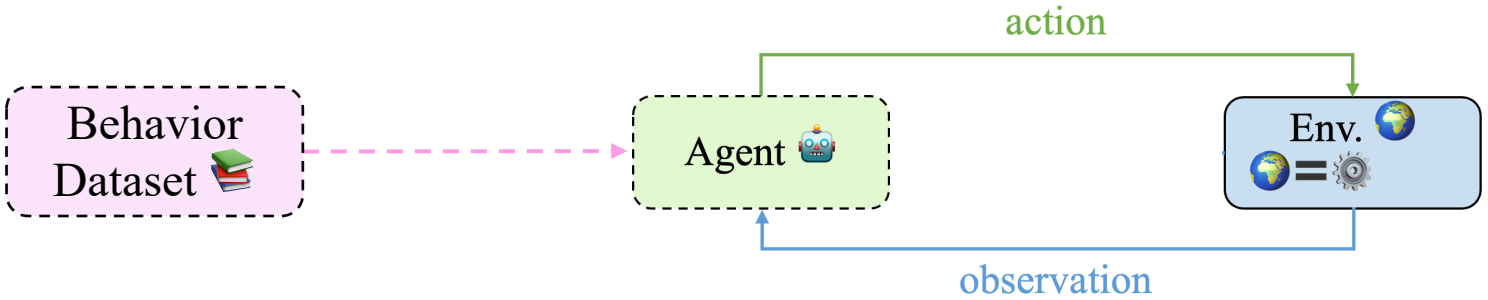
Methods for Learning from Behavior

- Learning from a *behavior dataset*
  - Imitation Learning: recover π^* given behavior generated by π^*

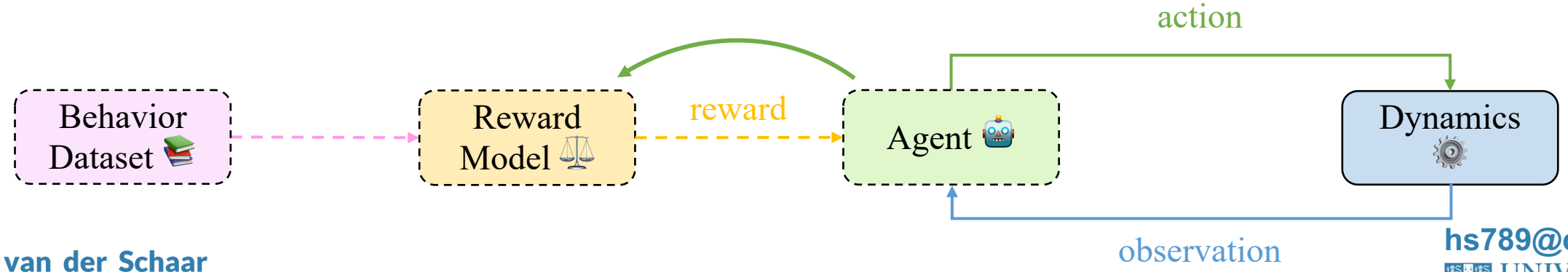


Methods for Learning from Behavior

- Learning from a *behavior dataset*
  - Imitation Learning: recover π^* given behavior generated by π^*



  - Inverse RL: recover R that induces π^* given behavior generated by π^*



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Imitation Learning

- Imitation Learning: recover π^* given behavior generated by π^*

  - IL 1. Behavior Clone
    [Hayes & Demiris, 1994] [Pomerleau, 1988]

Imitation Learning

- Imitation Learning: recover π^* given behavior generated by π^*

  - IL 1. Behavior Clone
    [Hayes & Demiris, 1994] [Pomerleau, 1988]

$$(s, a) \sim \mathcal{D} \quad \min_f |a - f(s)|$$

conditional action distribution matching

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Imitation Learning

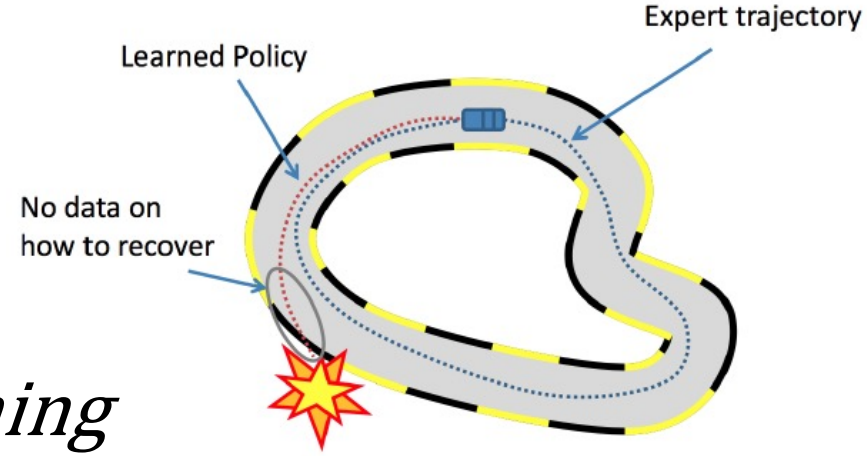
- Imitation Learning: recover π^* given behavior generated by π^*

- IL 1. Behavior Clone "*Compounding Error*" [Hayes & Demiris, 1994] [Pomerleau, 1988]

$$(s, a) \sim \mathcal{D} \quad \min_f |a - f(s)|$$

conditional action distribution matching

~~*state distribution matching*~~

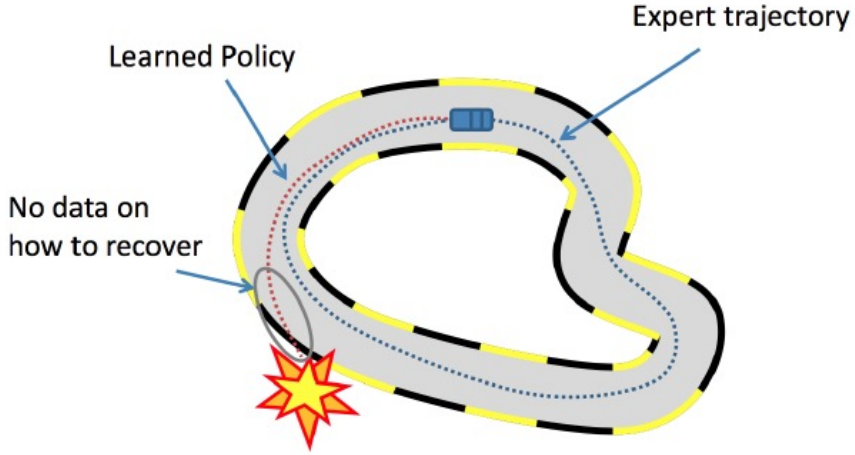


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Imitation Learning

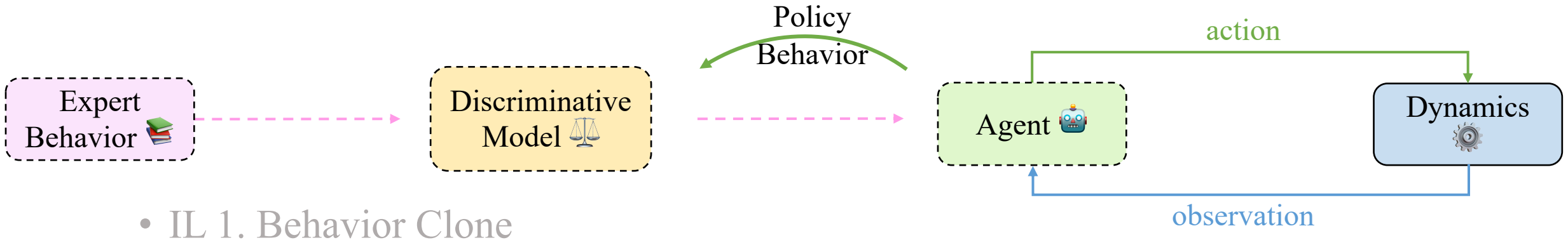
- Imitation Learning: recover π^* given behavior generated by π^*

- IL 1. Behavior Clone
  [Hayes & Demiris, 1994; Pomerleau, 1988]
- IL 2. Expert Engagement
  [Ross et al. 2011; Li et al., 2022; Peng et al. 2023]



Imitation Learning

- Imitation Learning: recover π^* given behavior generated by π^*

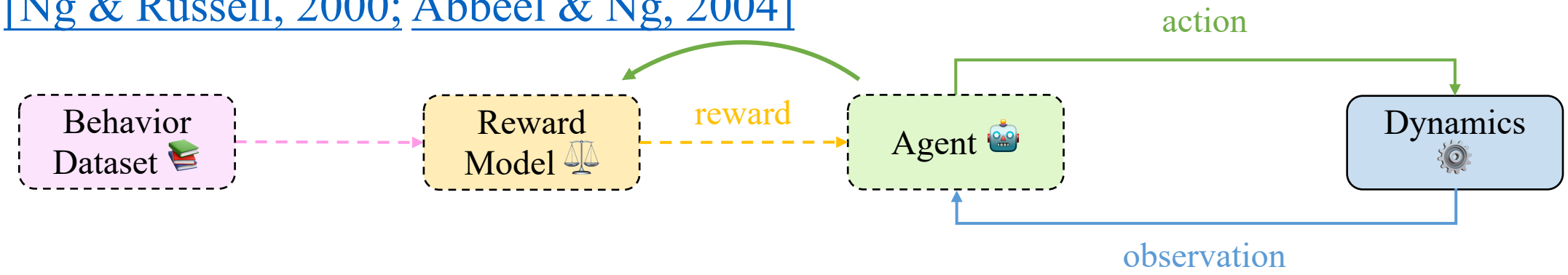


- IL 1. Behavior Clone
  [Hayes & Demiris, 1994; Pomerleau, 1988]
- IL 2. Expert Engagement
  [Ross et al. 2011; Li et al., 2022; Peng et al. 2023]
- IL 3. Adversarial Imitation (GAIL)
  [Ho & Ermon, 2016]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Inverse Reinforcement Learning

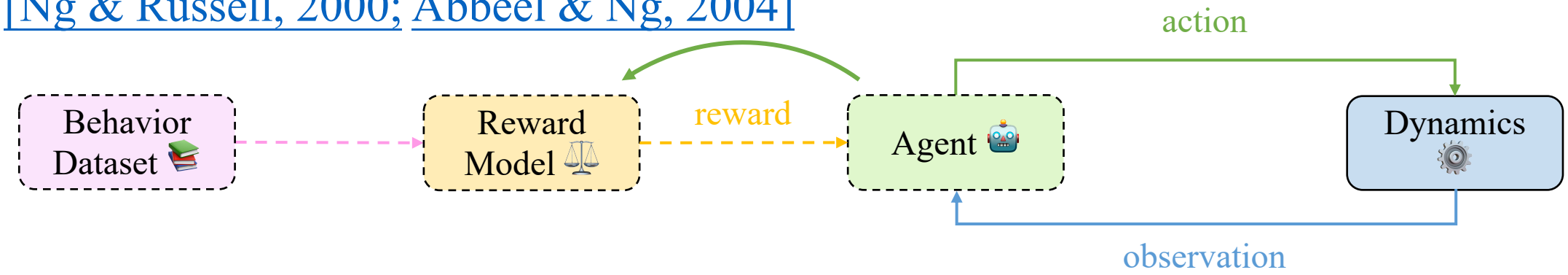
- Inverse RL: recover R that induces π^* given behavior generated by π^*
  [Ng & Russell, 2000; Abbeel & Ng, 2004]



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Inverse Reinforcement Learning

- Inverse RL: recover R that induces π^* given behavior generated by π^*
  [Ng & Russell, 2000; Abbeel & Ng, 2004]



$\text{MDP}\backslash R + \hat{R} \rightarrow \text{MDP}$

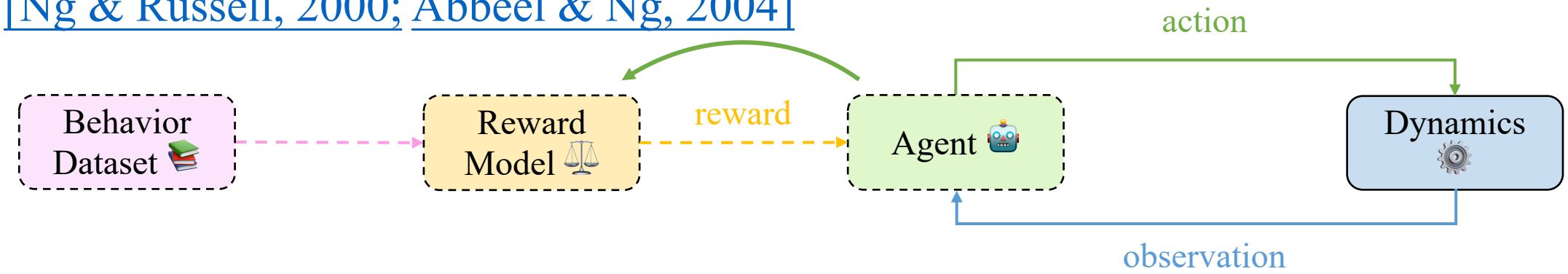
~~$\text{MDP}\backslash R$:~~ MDP missing ~~Reward Function~~

$\hat{R}$: Reward Model (from behavior data)

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Inverse Reinforcement Learning

- Inverse RL: recover R that induces π^* given behavior generated by π^* [Ng & Russell, 2000; Abbeel & Ng, 2004]

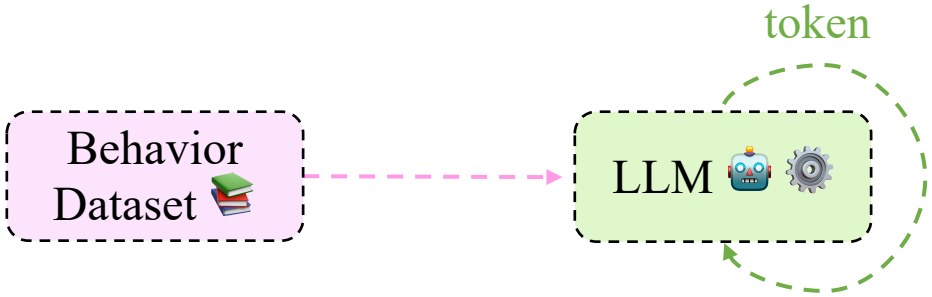


  - Max-Ent IRL [Ziebart et al., 2008]
  - Adversarial IRL [Fu et al., 2017]
  - T-REX [Brown et al., 2019]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Optimization via Imitation

- LLMs as language imitators

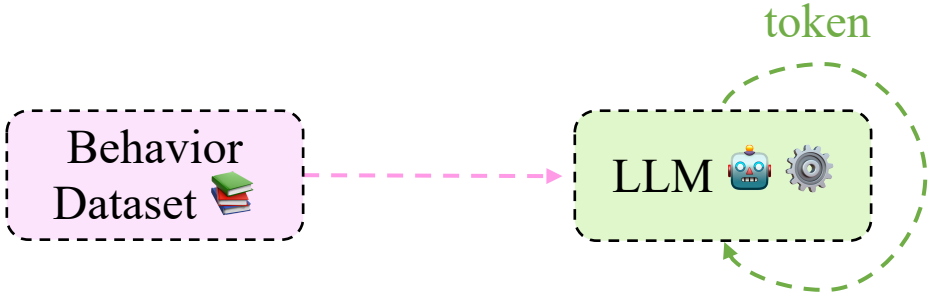


- Pre-train: large scale behavior clone
  [Obtain (strong) ability of understanding]

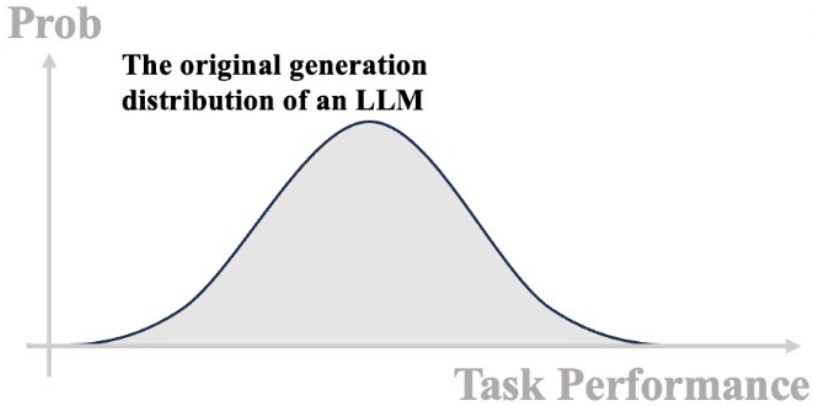
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Optimization via Imitation

- LLMs as language imitators



- Pre-train: large scale behavior clone
  [Obtain (strong) ability of understanding]

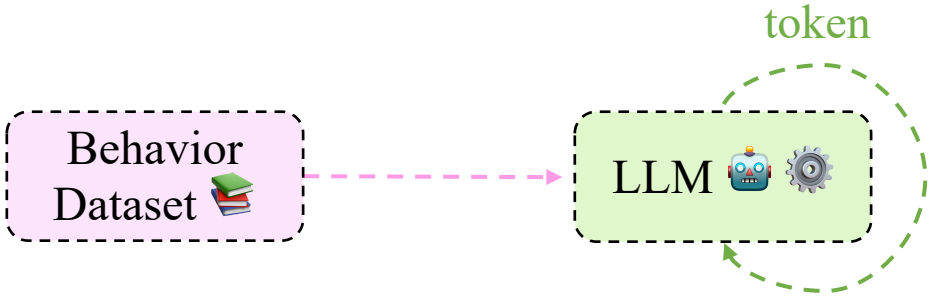


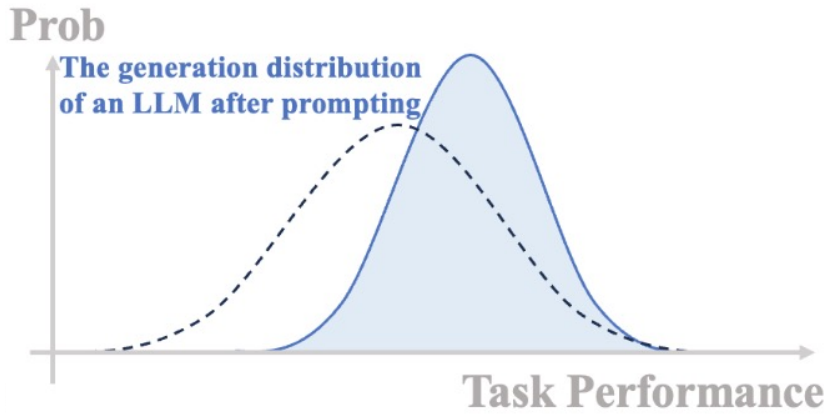
(1) LLM *Can* do Any Task as a *Universal Sampler*

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Optimization via Imitation

- LLMs as language imitators





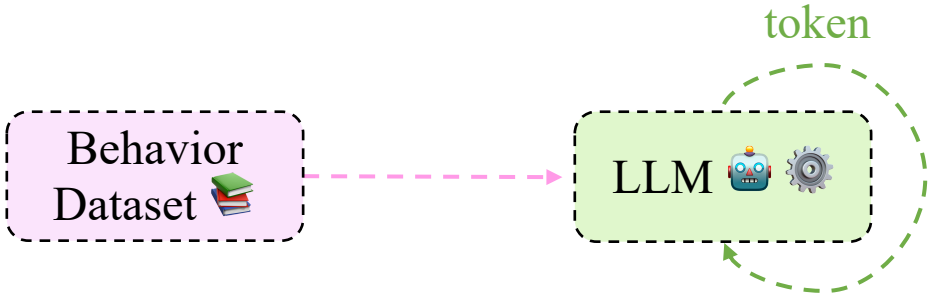
(2) Prompting *Can* Improve Performance by shifting the generation

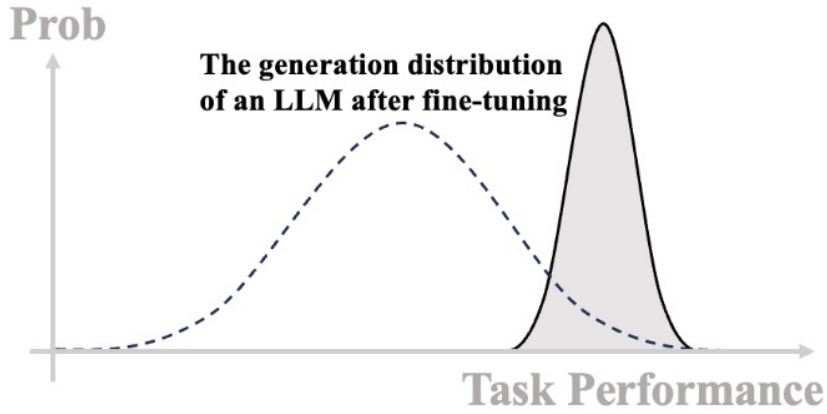
- Pre-train: large scale behavior clone
  [Obtain (strong) ability of understanding]
- Post-train/alignment: optimization on a specific task
  - Smart prompting strategy [Kojima et al., 2022]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Optimization via Imitation

- LLMs as language imitators





(3) Fine-Tuning *Can* Improve Performance by shifting the generation

- Pre-train: large scale behavior clone
  [Obtain (strong) ability of understanding]
- Post-train/alignment: optimization on a specific task
  - Smart prompting strategy [Kojima et al., 2022]
  - Supervised fine-tuning

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Why Do We Need Inverse RL?

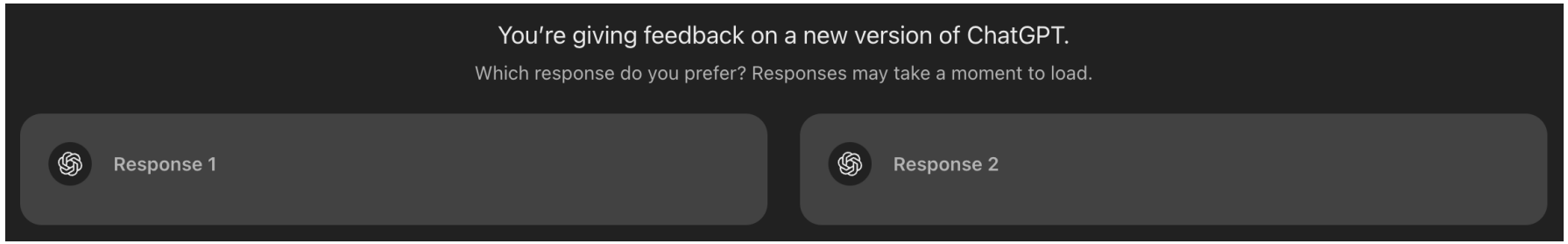
-

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Why Do We Need Inverse RL?

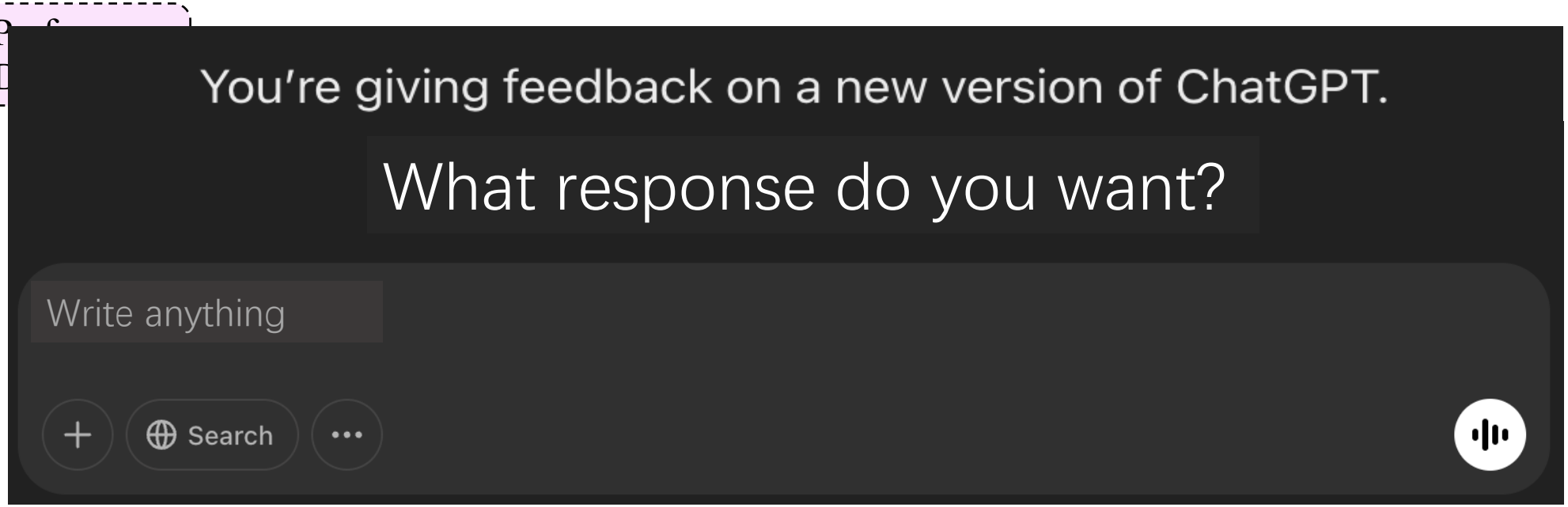
- Preference feedback can be more scalable & practical

Preference Dataset 📚



Why Do We Need Inverse RL?

- Preference feedback can be more scalable & practical

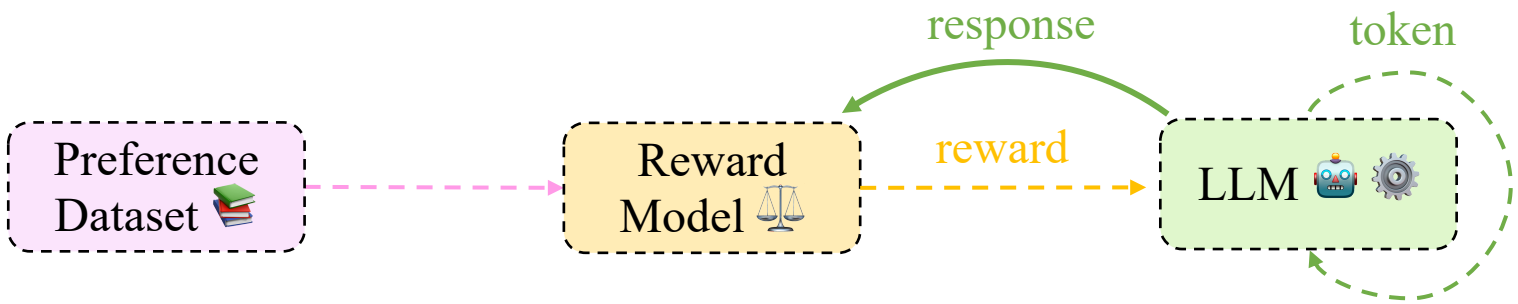


(this is not real)

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

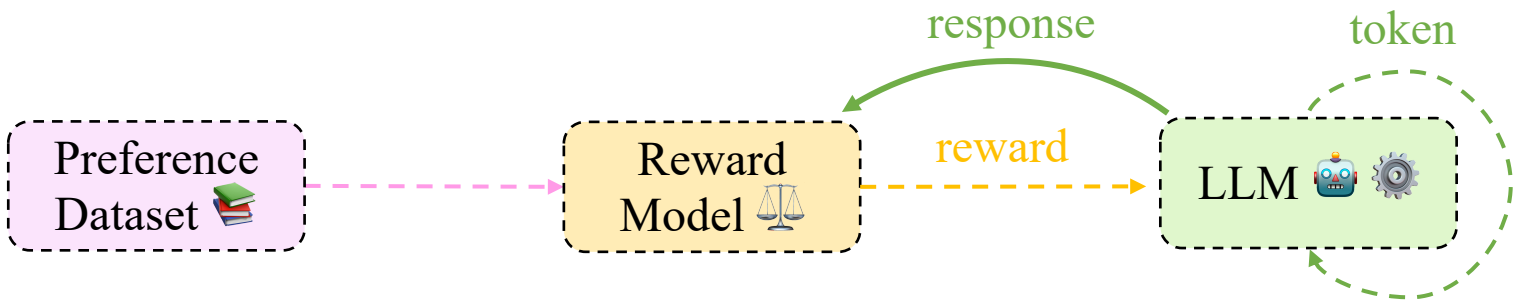
Why Do We Need Inverse RL?

- Preference feedback can be more scalable & practical [RLHF]

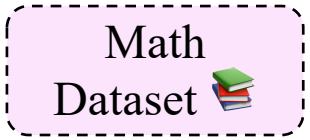


Why Do We Need Inverse RL?

- Preference feedback can be more scalable & practical [RLHF]



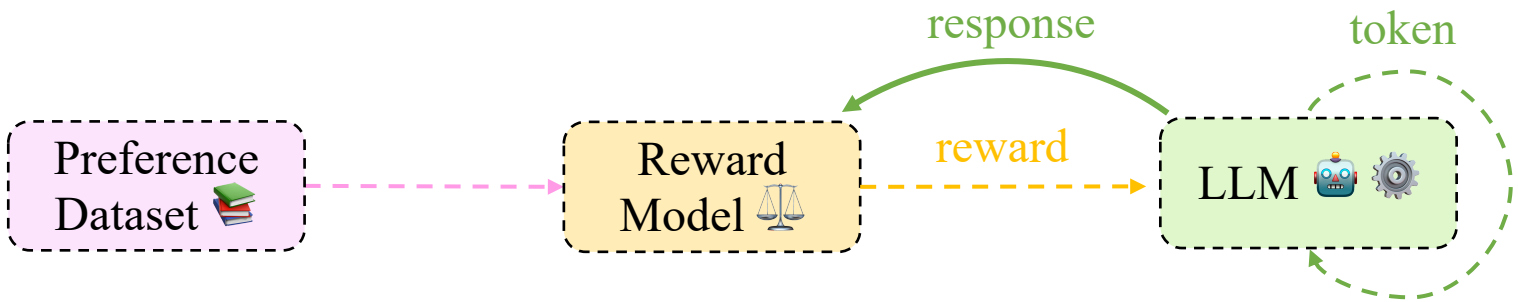
- Math: find a more generalizable reasoning path toward correct answers



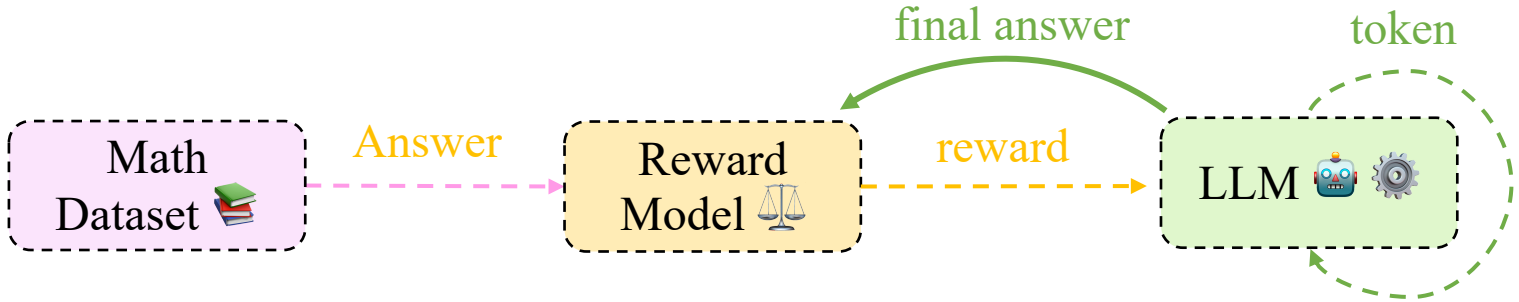
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Why Do We Need Inverse RL?

- Preference feedback can be more scalable & practical [RLHF]

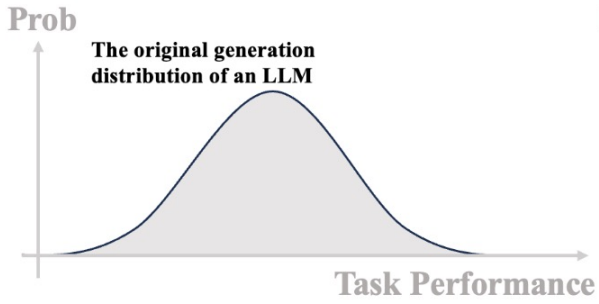


- Math: find a more generalizable reasoning path toward correct answers

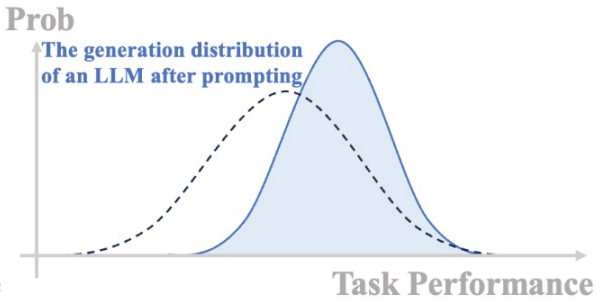


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

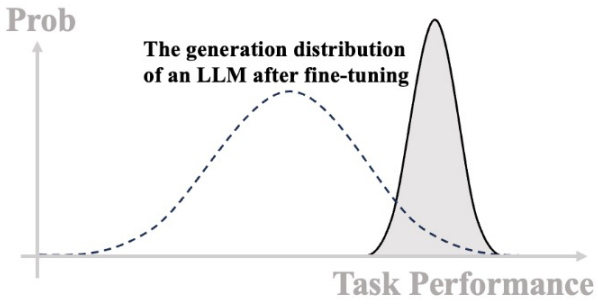
RMs Enable Test-Time Optimization [Sun et al., 2024]



(1) LLM *Can* do Any Task as a *Universal Sampler*



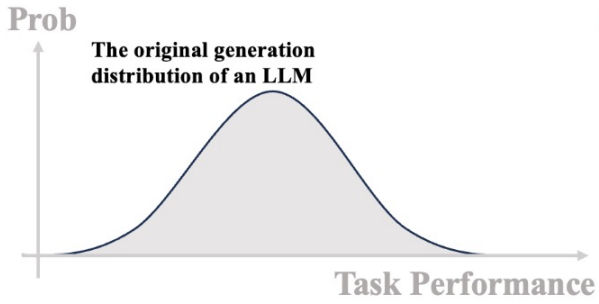
(2) Prompting *Can* Improve Performance by shifting the generation



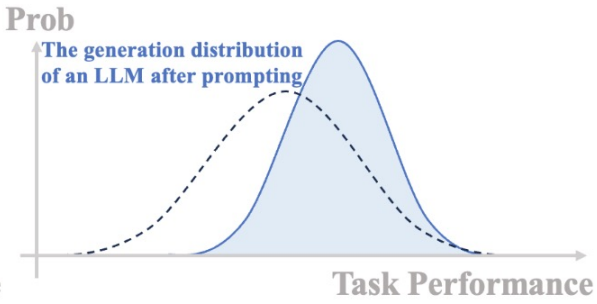
(3) Fine-Tuning *Can* Improve Performance by shifting the generation

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

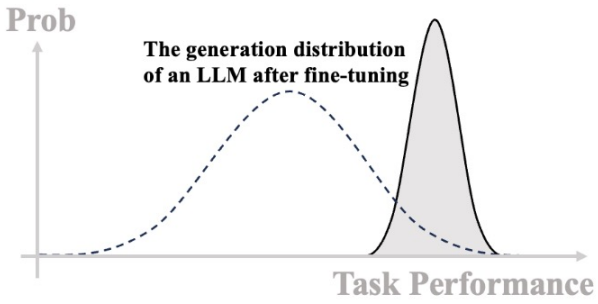
RMs Enable Test-Time Optimization [Sun et al., 2024]



(1) LLM *Can* do Any Task as a *Universal Sampler*



(2) Prompting *Can* Improve Performance by shifting the generation



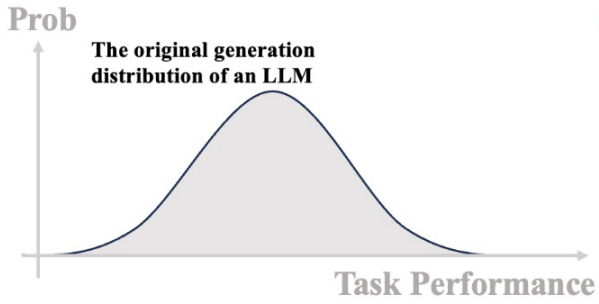
(3) Fine-Tuning *Can* Improve Performance by shifting the generation

Train on Test Task

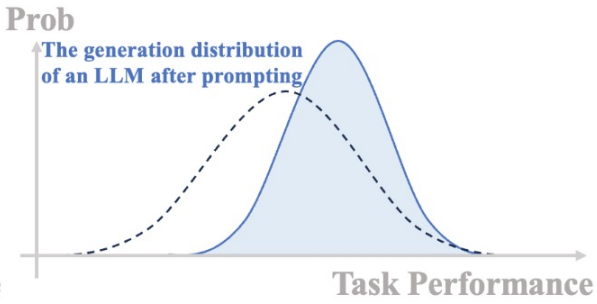
One forward pass is enough

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

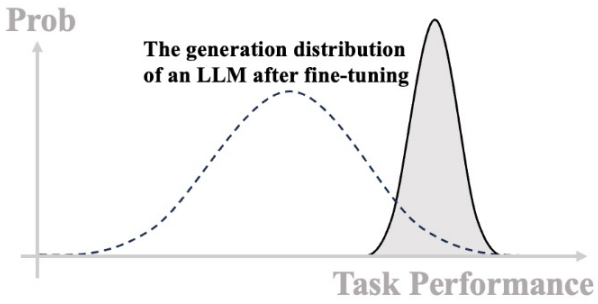
RMs Enable Test-Time Optimization [Sun et al., 2024]



(1) LLM *Can* do Any Task as a *Universal Sampler*



(2) Prompting *Can* Improve Performance by shifting the generation



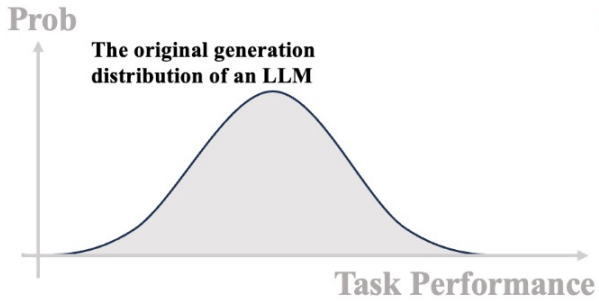
(3) Fine-Tuning *Can* Improve Performance by shifting the generation

Train on Test Task

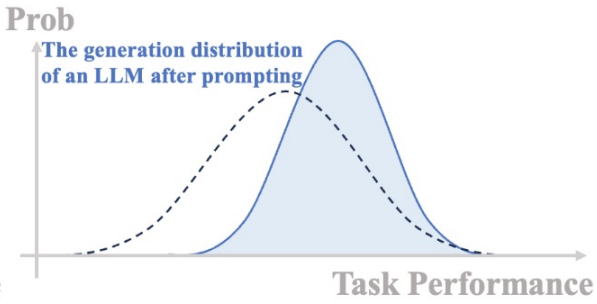
But one forward pass is *not* enough

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

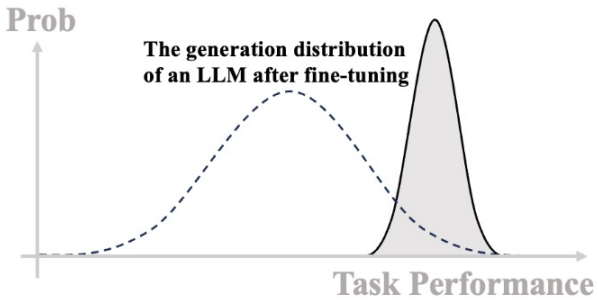
RMs Enable Test-Time Optimization [Sun et al., 2024]



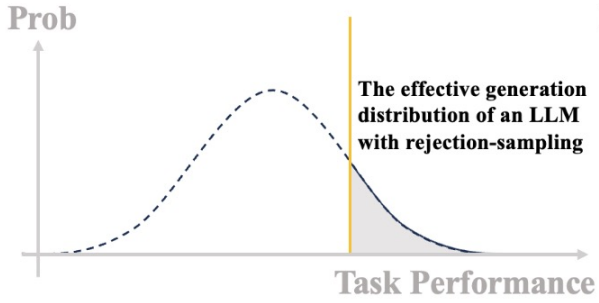
(1) LLM *Can* do Any Task as a *Universal Sampler*



(2) Prompting *Can* Improve Performance by shifting the generation



(3) Fine-Tuning *Can* Improve Performance by shifting the generation

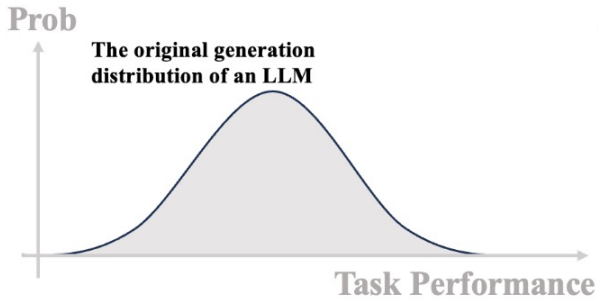


(4) Rejection-Sampling with *Reward Models* *Can* Improve Performance by filtering the generation

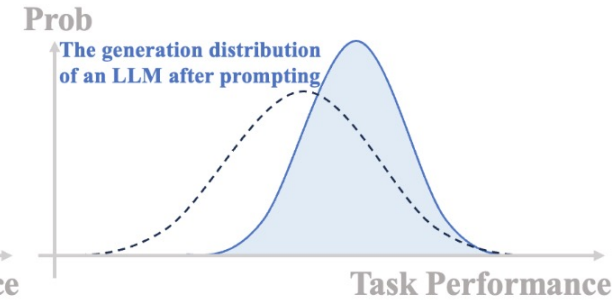
Reward Models
Enable Test-time Optimization

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

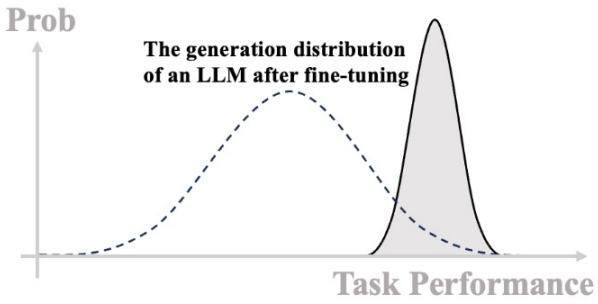
RMs Enable Test-Time Optimization [Sun et al., 2024]



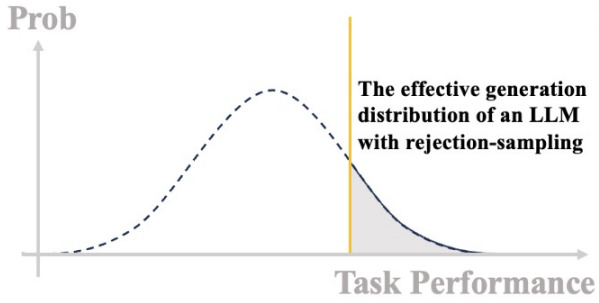
(1) LLM *Can* do Any Task as a *Universal Sampler*



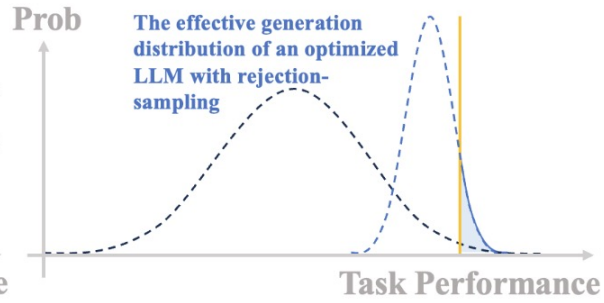
(2) Prompting *Can* Improve Performance by shifting the generation



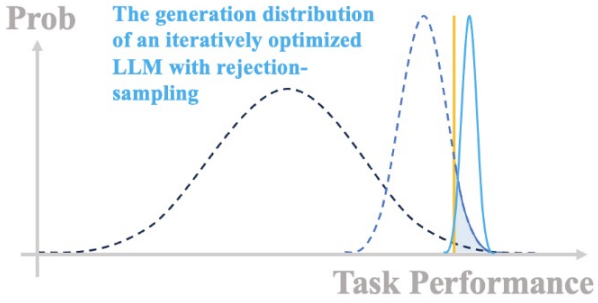
(3) Fine-Tuning *Can* Improve Performance by shifting the generation



(4) Rejection-Sampling with *Reward Models* *Can* Improve Performance by filtering the generation



(5) On Hard Tasks, Reward Models are Crucial as they enable search and *Inference-Time-Optimization*



(6) Searching with Reward Models can generate datasets that enable *iterative fine-tuning*

Reward Models Enable Test-time Optimization

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Takeaways:

- 1. RL can be formally described as MDP
- 2. There is no silver bullet in RL
- 3. IL and Inverse RL as MDP\R: policy learning from behavior
- 4. LLM pre-train/ SFT are imitation learning
- 5. Reason use Inverse RL:
  - Scalable & Flexible
  - Generalizes better than SFT
  - Enables Test-Time Optimization

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Part 3:

Learning Reward Models from Data

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Learning Reward Models from Data

IRL for Conversational AI (e.g., Classical RLHF)

IRL for Math Reasoning

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Learning Reward Models from Data

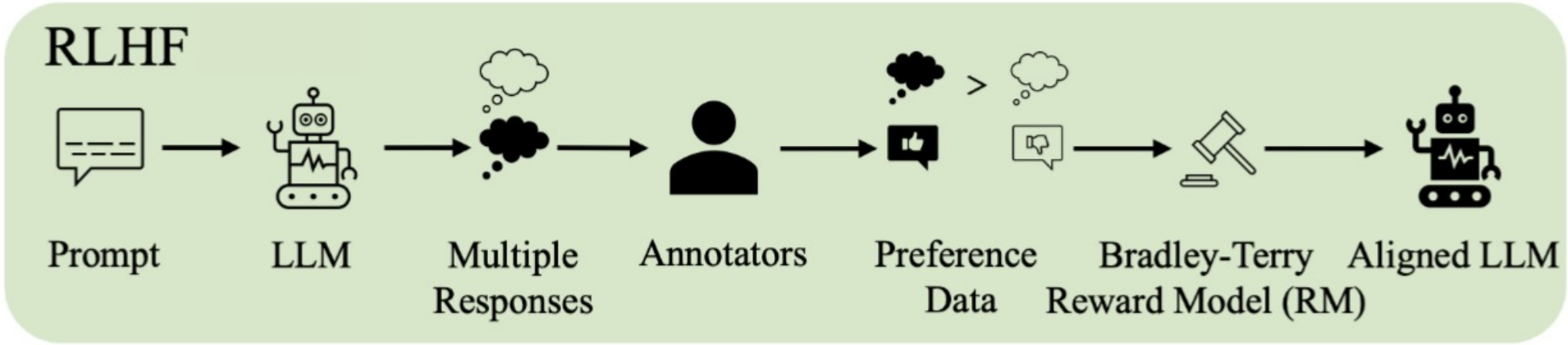
IRL for Conversational AI (e.g., Classical RLHF)

IRL for Math Reasoning

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

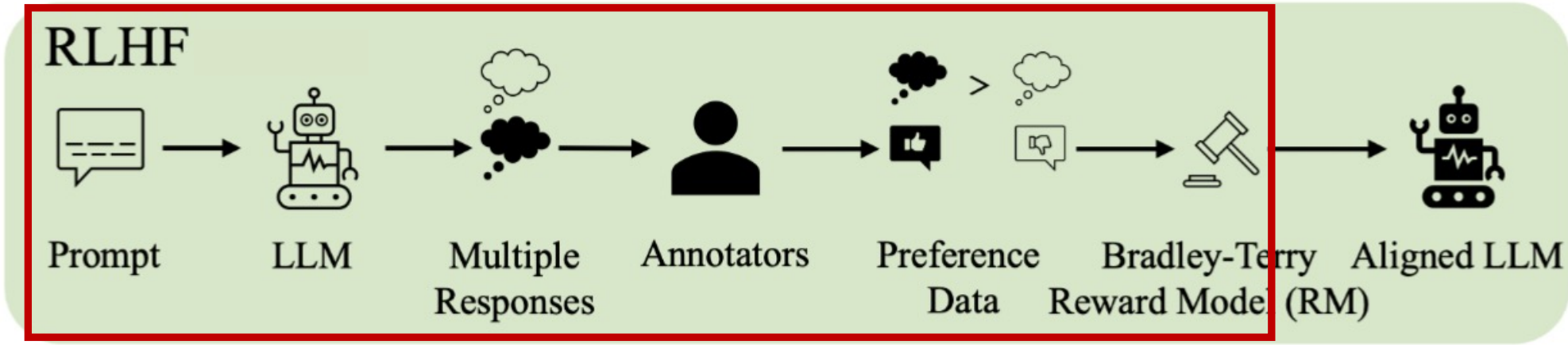
- RLHF as Inverse RL: Reward Modeling + forward RL



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- RLHF as Inverse RL: Reward Modeling + forward RL

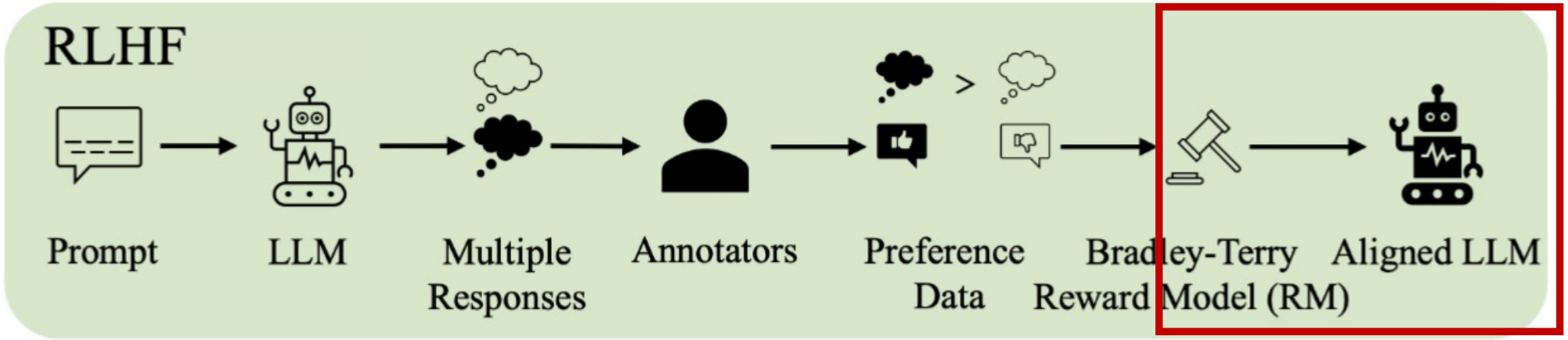


Reward Modeling

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- RLHF as Inverse RL: Reward Modeling + forward RL

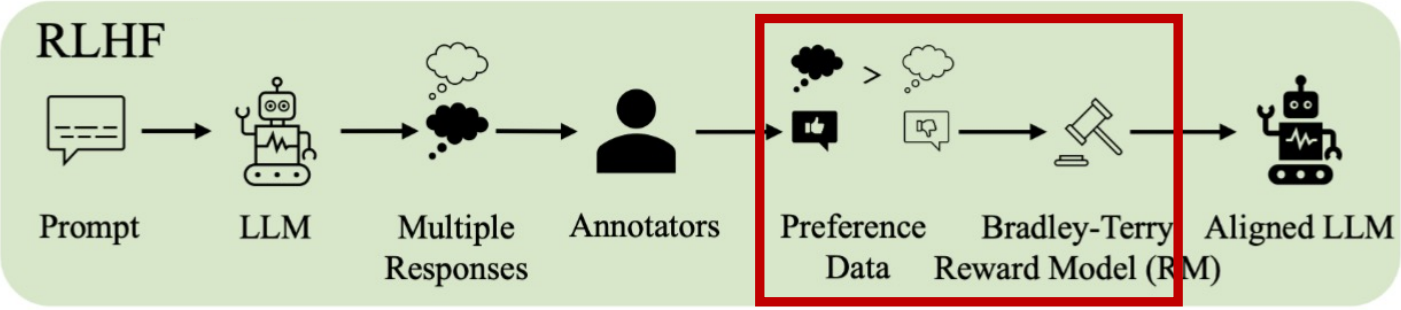


Policy Learning

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RLHF with *Bradley-Terry* Reward Models

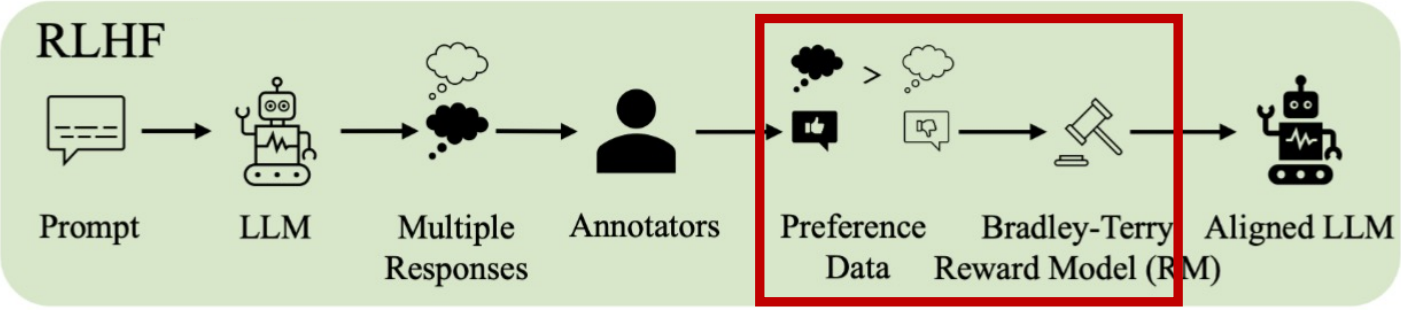
- How to build reward model?
  - "use the Bradley-Terry Model"



  - But ***why***?

RLHF with *Bradley-Terry* Reward Models

- How to build reward model?
  - "use the Bradley-Terry Model"



- What is the Bradley Terry Model?
  - Player i, with ability score r_i
  - Player j, with ability score r_j
  - In a game between player i and j,

$$P(i \text{ wins } j) = \frac{r_i}{r_i + r_j}$$

Estimate BT Score: Parameter Estimation

- Classical Applications of the Bradley-Terry Models
  - In Chess/StarCraft/DOTA II: we estimate player scores using match history/outcomes
  - In LLM arena, we have ~200 models, 2M competitions, each LLM plays ~10,000 games

Estimate BT Score: Parameter Estimation

- Classical Applications of the Bradley-Terry Models
  - In Chess/StarCraft/DOTA II: we estimate player scores using match history/outcomes
  - In LLM arena, we have ~200 models, 2M competitions, each LLM plays ~10,000 games

Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization
1	Gemini-2.5-Pro-Exp-03-25	1439	+7/-10	5858	Google
1	ChatGPT-4o-latest (2025-03-26)	1410	+8/-10	4899	OpenAI
2	GPT-4.5-Preview	1398	+5/-7	12312	OpenAI
4	Grok-3-Preview-02-24	1403	+6/-6	12391	xAI
4	Gemini-2.0-Pro-Exp-02-05	1380	+4/-4	20289	Google
4	DeepSeek-V3-0324	1369	+10/-10	3526	DeepSeek
4	o1-2024-12-17	1351	+5/-4	26722	OpenAI

Image source: lmsys-llm-arena, screenshot date: April-7-2025

$N\log(N)$: Average Complexity for Sorting

- Classical Applications of the Bradley-Terry Models --- Parameter Estimation
  - In Chess/StarCraft/DOTA II: we estimate player scores using match history/outcomes
  - In LLM arena, we have ~200 models, 2M competitions, each LLM plays ~10,000 games

We need a **large number of matches/games** for a consistent estimation.
e.g., consider *sorting*: we need $\boldsymbol{O(N\log N)}$

Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

RLHF: *Prediction* w/ $\ll N\log(N)$ Sample

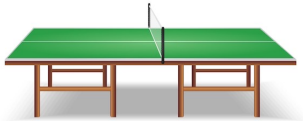
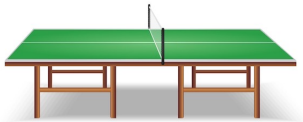
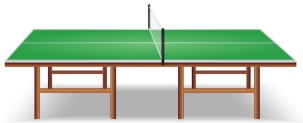
- Classical Applications of the Bradley-Terry Models --- Parameter Estimation
  - In Chess/StarCraft/DOTA II: we estimate player scores using match history/outcomes
  - In LLM arena, we have ~200 models, 2M competitions, each LLM plays ~10,000 games

  We need a **large number of matches/games** for a consistent estimation.
  e.g., consider *sorting*: we need $\boldsymbol{O(N\log N)}$

- In RLHF,
  - we have $\boldsymbol{K}$ prompts, $\boldsymbol{2K}$ responses --- $\boldsymbol{2K}$ players
  - Each pair only "compete" once --- $\boldsymbol{K \ll 2K\log(2K)}$ comparisons
  - + we need predictions, how is this possible?

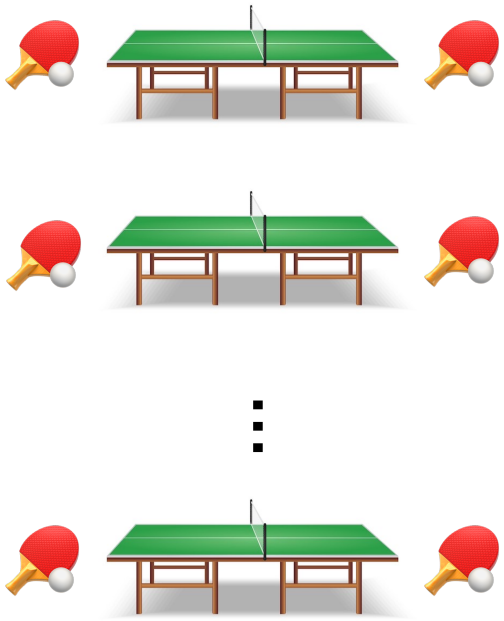
BT in RLHF: Mission Impossible?

- Why this is challenging (Why does BT model work?)
- Analogy
  - K Prompts – $2K$ Responses – K Comparisons – Predict Best of N (test time)
  - K Tables – $2K$ Players – K Games – ***Predict*** Best of N ***New*** Player ?

BT in RLHF: Mission Impossible?

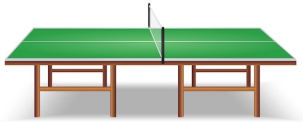
- Why this is challenging (Why does BT model work?)
- Analogy
  - K Prompts – $2K$ Responses – K Comparisons – Predict Best of N (test time)
  - K Tables – $2K$ Players – K Games – ***Predict*** Best of N ***New*** Player ?



Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

BT in RLHF: Mission Impossible?

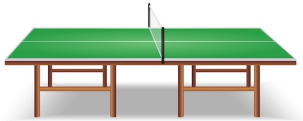
- Why this is challenging (Why does BT model work?)
- Analogy
  - K Prompts – $2K$ Responses – K Comparisons – Predict Best of N (test time)
  - K Tables – $2K$ Players – K Games – ***Predict*** Best of N ***New*** Player ?



  - This is impossible, unless …

Mission ~~im~~Possible w/ *Good Features*

- Why this is challenging (Why does BT model work?)
- Analogy
  - K Prompts – $2K$ Responses – K Comparisons – Predict Best of N (test time)
  - K Tables – $2K$ Players – K Games – ***Predict*** Best of N ***New*** Player ?



  - This is impossible, unless we do regression w/ **generalizable features/covariates** (e.g., years played, equipment prices, formal training)

Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

RLHF w/ BT: *Embeddings Generalize*

- Why does BT model work?
  - We are working on the ***embedding space***
  - RMs in the embedding space generalize well
  - Theoretical justification is in the paper

Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

Is BT model Necessary in RLHF?

- Why does BT model work?
  - We are working on the *embedding space*
  - RMs in the embedding space generalize well
  - Theoretical justification is in the paper

- Is BT model necessary?

Is BT model Necessary in RLHF?

- Why does BT model work?
  - We are working on the *embedding space*
  - RMs in the embedding space generalize well
  - Theoretical justification is in the paper

- Is BT model necessary?
  - BT model: precisely predicting win rates (make a bet, match players)

Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

~~Win Rate?~~ Order Consistency!

- Why does BT model work?
  - We are working on the *embedding space*
  - RMs in the embedding space generalize well
  - Theoretical justification is in the paper

- Is BT model necessary?
  - BT model: precisely predicting win rates
  - RLHF?

NO, we only need ***order consistency***

Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

~~Win Rate?~~ Order Consistency!

- Why does BT model work?
  - We are working on the *embedding space*
  - RMs in the embedding space generalize well
  - Theoretical justification is in the paper

- Is BT model necessary?
  - BT model: precisely predicting win rates
  - RLHF?

    NO, we only need ***order consistency***

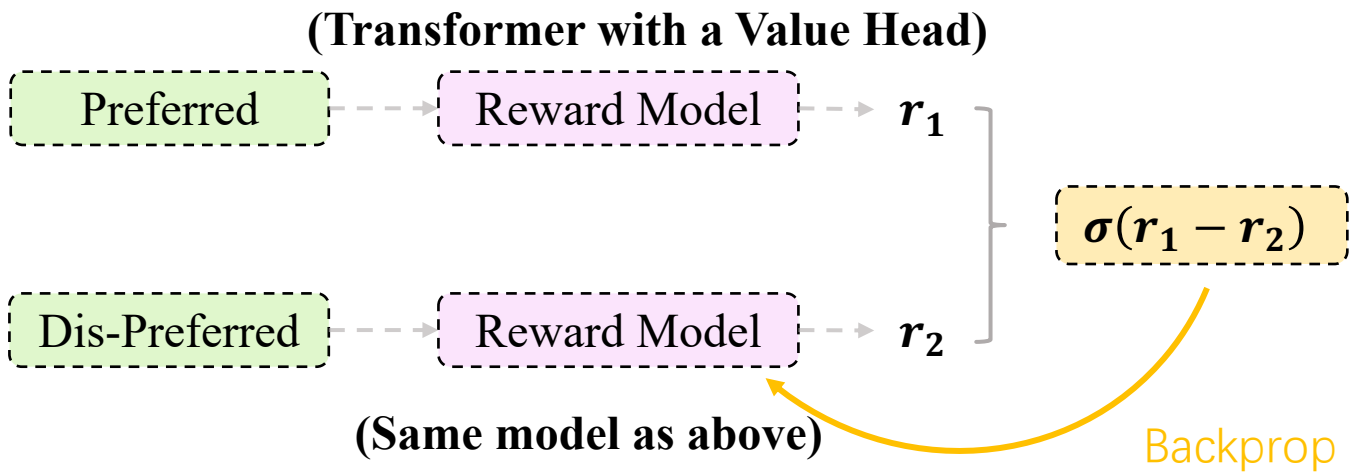
  - Classification RM --- marginalized win rates

Classification RM: Flexible

- Classification models are…
  - Flexible

	Input	Output	Model
Bradley Terry Model	A Pair	Score difference	Neural Networks
Classification Model	1 response	Score	Any ML Models

Bradley-Terry Reward Model

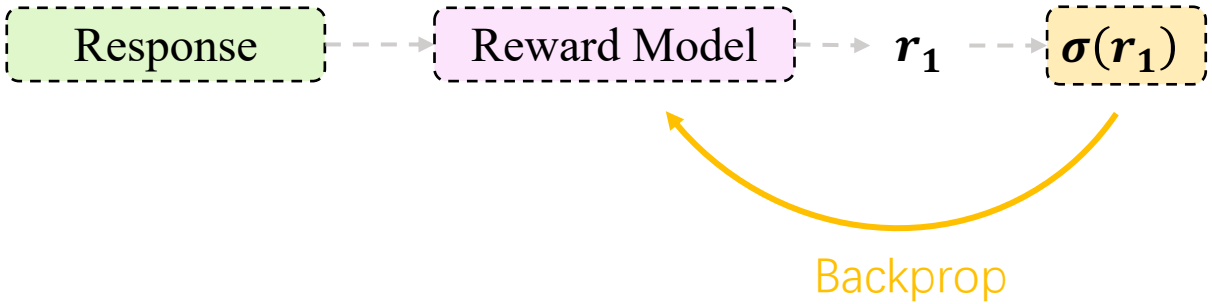


Classification RM: Flexible

- Classification models are…
  - Flexible

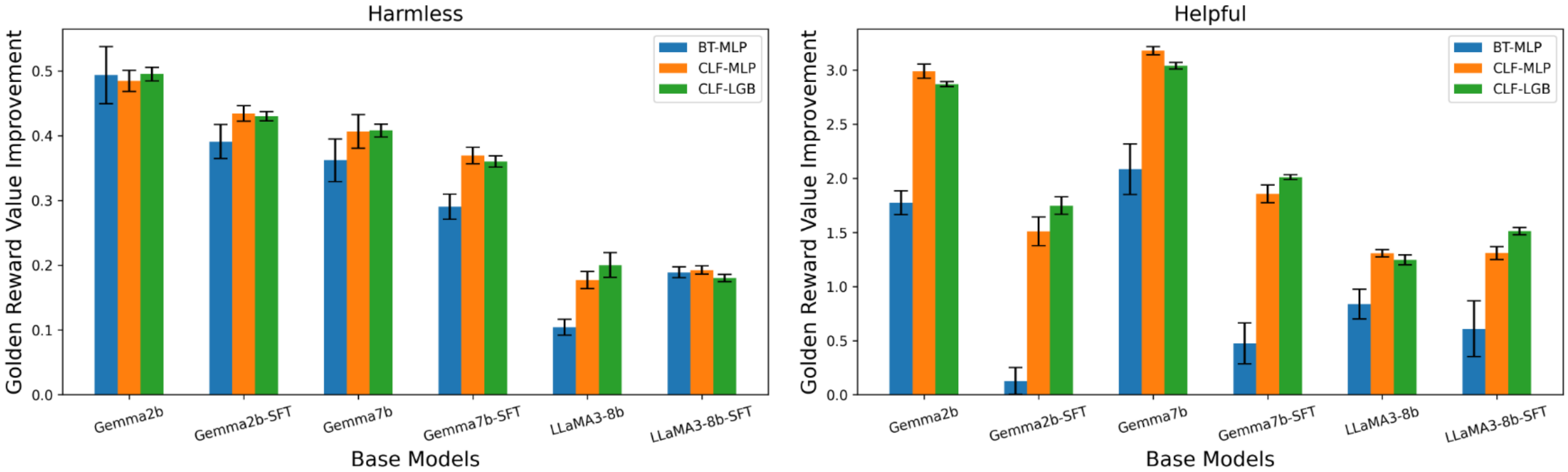
	Input	Output	Model
Bradley Terry Model	A Pair	Score difference	Neural Networks
Classification Model	1 response	Score	Any ML Models

Classification-based Reward Model



Clf RM: Better Performance than BT

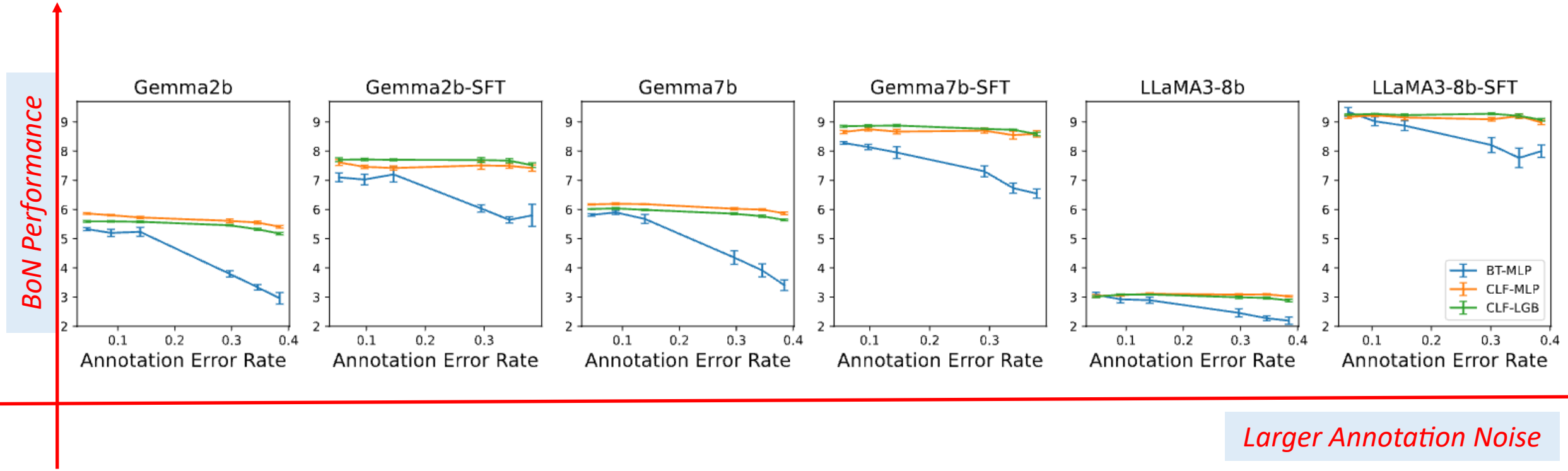
- Classification models are…
  - Flexible
  - Better than BT models (esp. when *labels are noisy*)



Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

Clf RM: More Robust to Annotation Noise

- Classification models are…
  - Flexible
  - Better than BT models (esp. when *labels are noisy*)
  - More robust to annotation noise



RM from Preference: Takeaways

- Preference learning with less than $N \log N$ comparisons relies on the embeddings.
- Bradley-Terry models are good, but not necessary.
- Order consistency is the objective of RM.
- Classification methods work well.

- Applications?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

RM from Cross-Prompt Comparisons

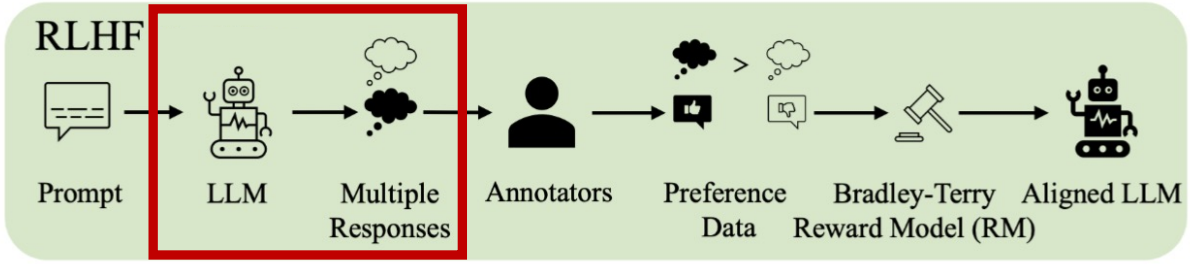
- Classical BT model applications: ***Randomized Comparisons***
- All embeddings are directly comparable

RM from Cross-Prompt Comparisons

- Classical BT model applications: ***Randomized Comparisons***
- All embeddings are directly comparable
- Example of *Cross-Prompt Comparison*:
  is the LLM more helpful in solving problem A than problem B

RM from Cross-Prompt Comparisons

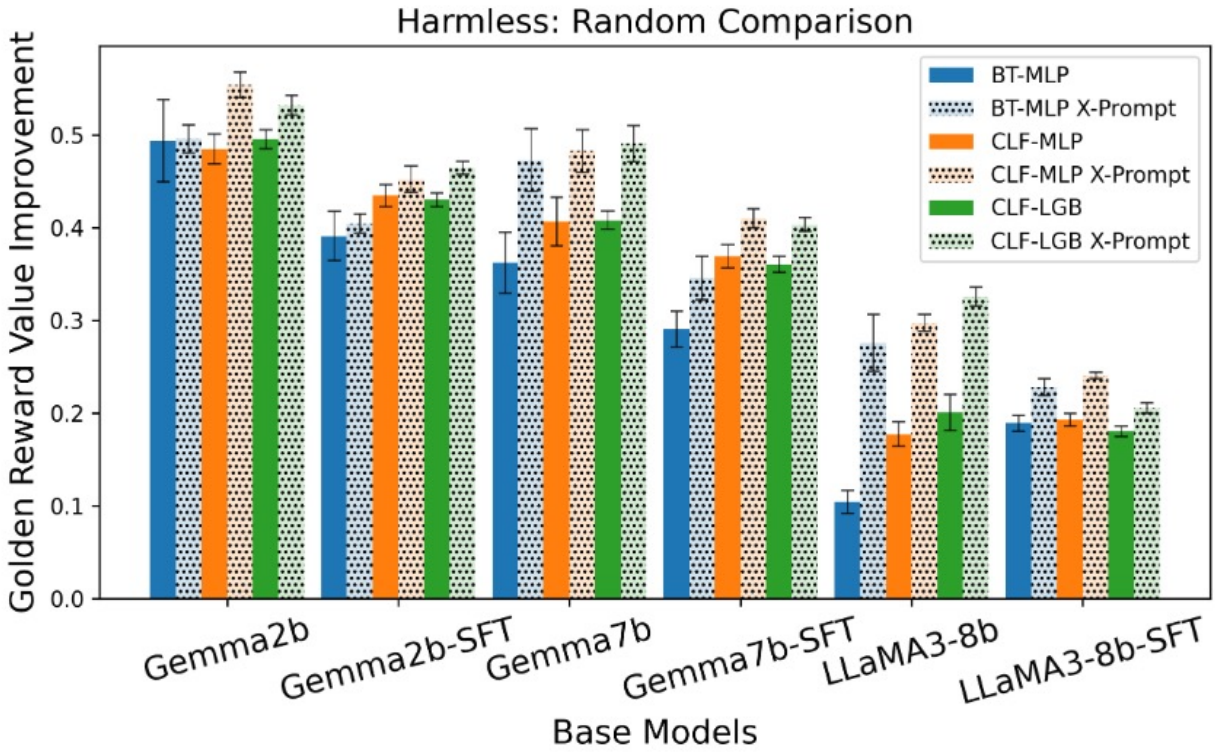
- Classical BT model applications: ***Randomized Comparisons***
- All embeddings are directly comparable
- Example of *Cross-Prompt Comparison*:
  is the LLM more helpful in solving problem A than problem B
- What if all responses are equally great?

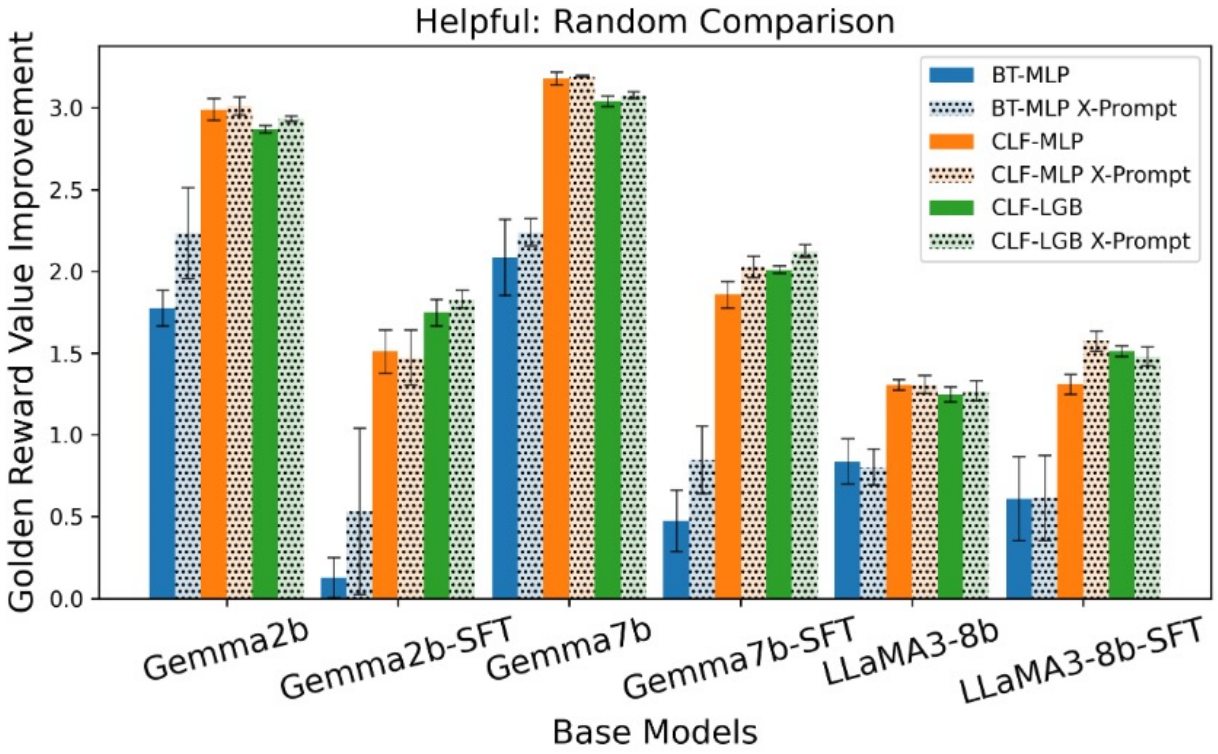


Rethinking Reward Modeling [Hao Sun*; Yunyi Shen*; Jean-François Ton, ICLR'2025]

RM from Cross-Prompt Comparisons

- Results: Cross-Prompt Comparisons lead to better reward models.

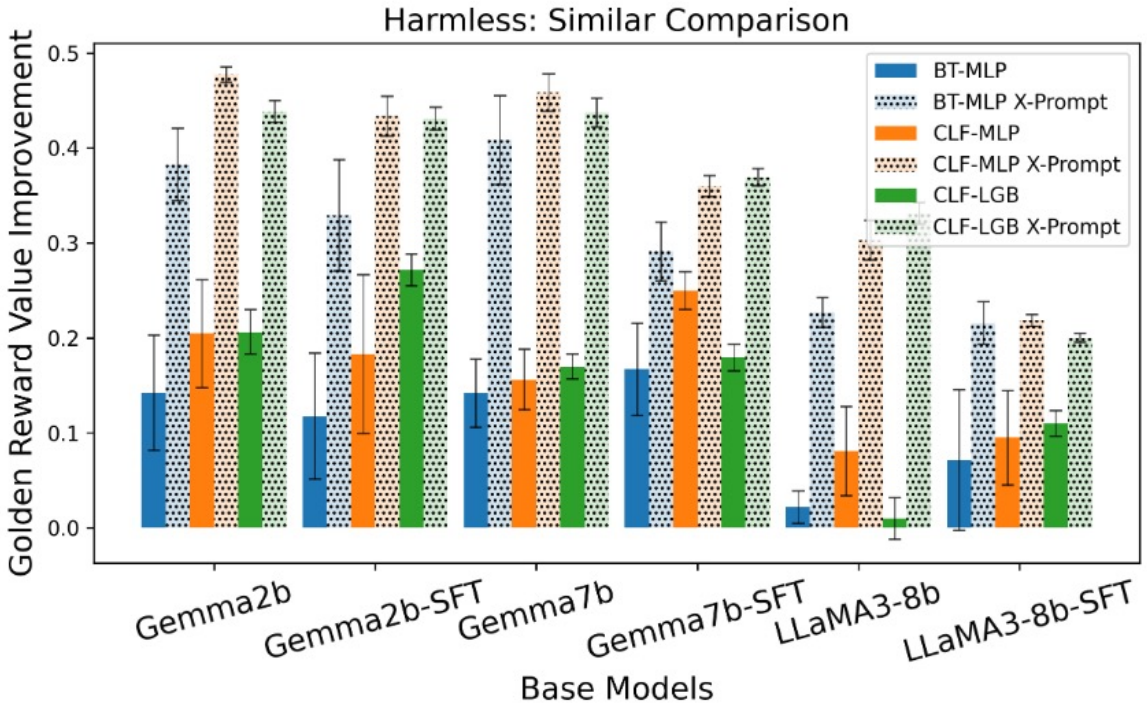


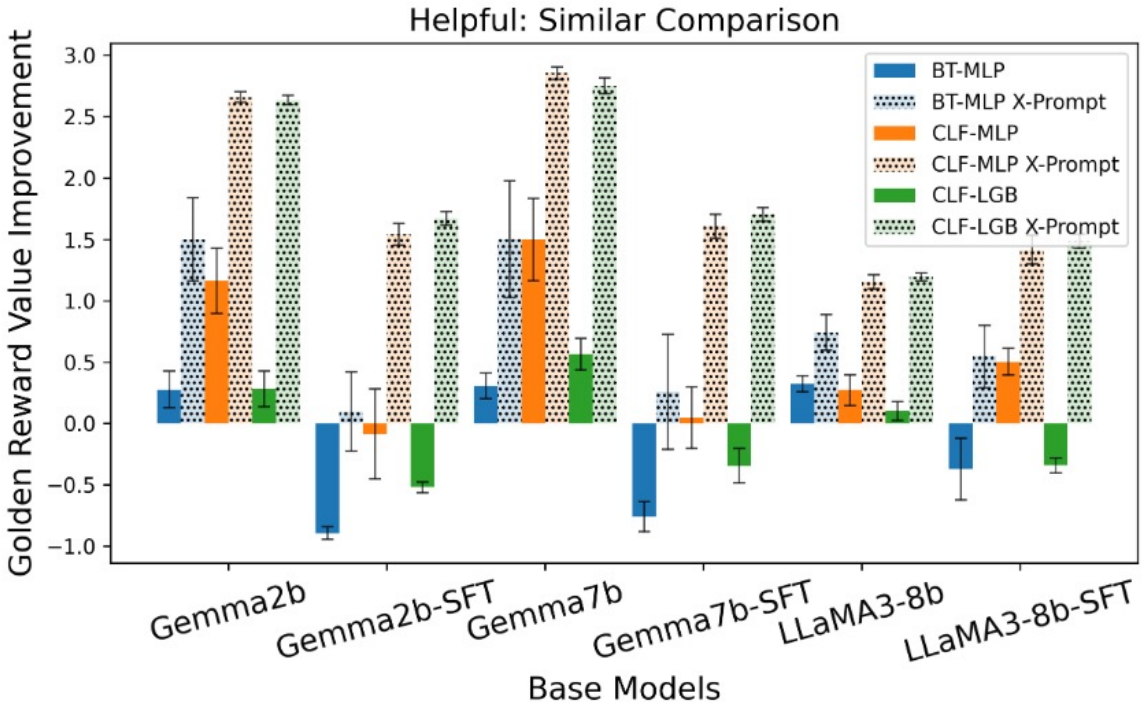


Cross-Prompt in Low Diversity

- When and Why? Diversity Matters

  Low diversity case: significant improvement

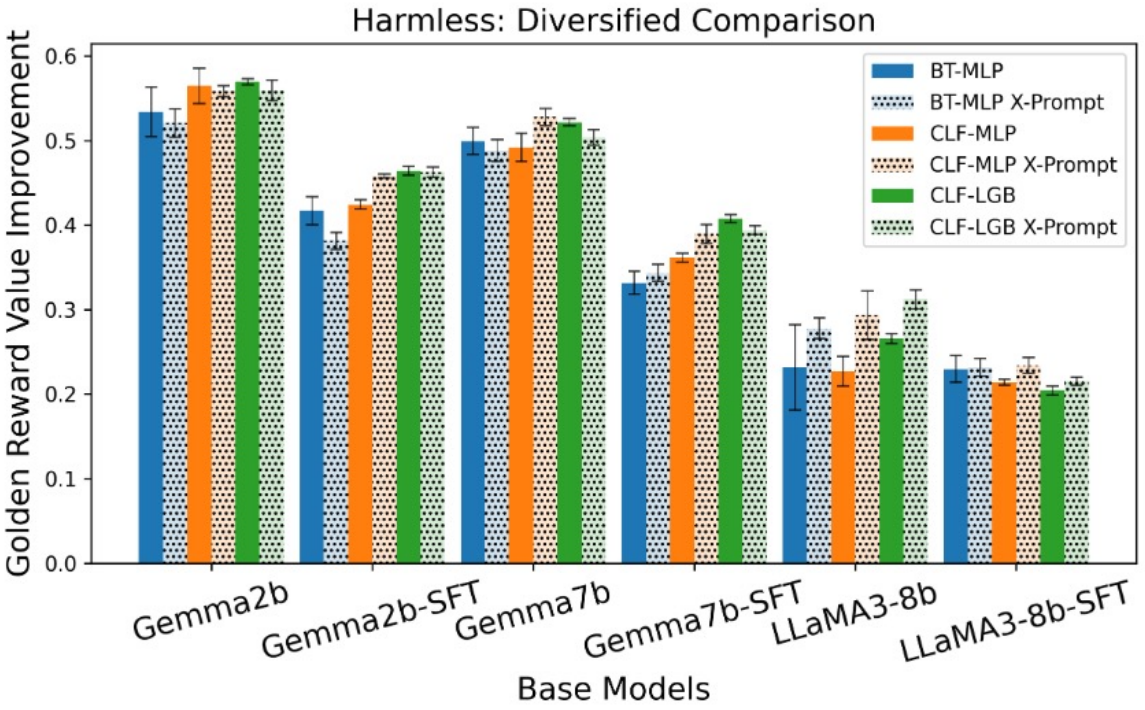


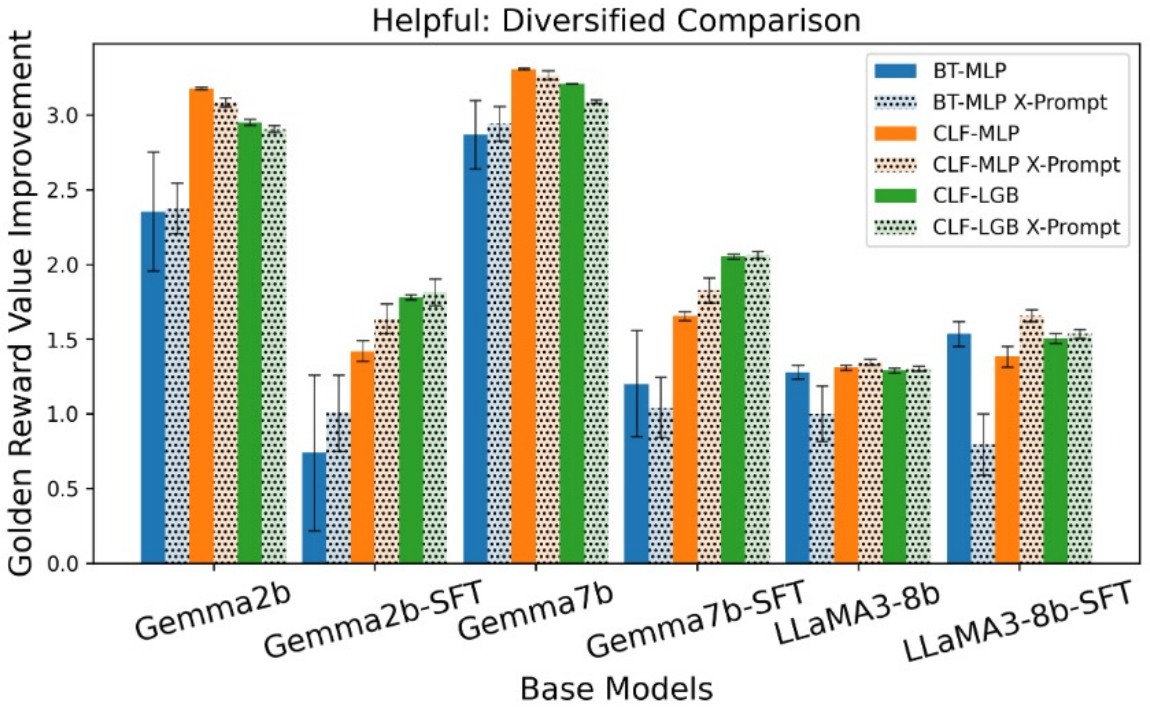


Cross-Prompt in High Diversity

- When and Why? Diversity Matters

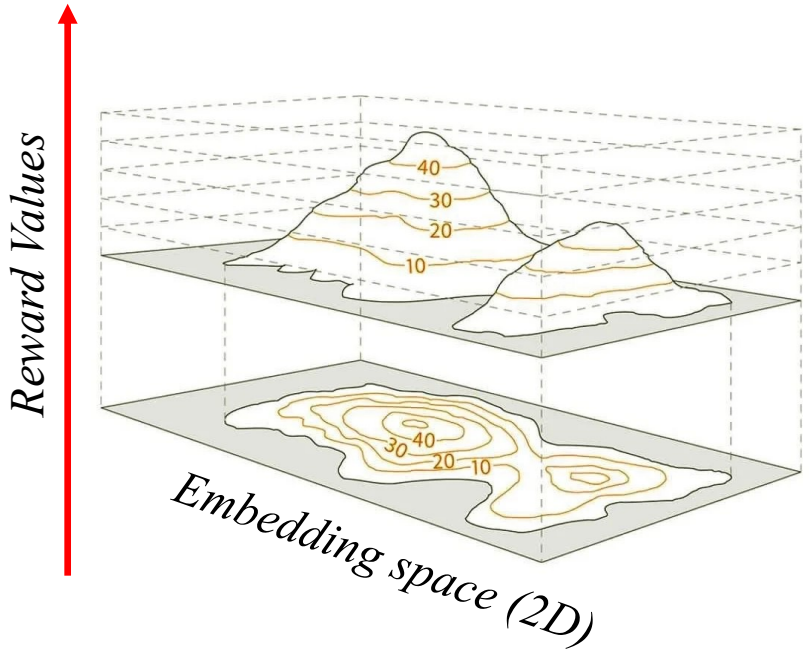
  High diversity case: negligible improvement





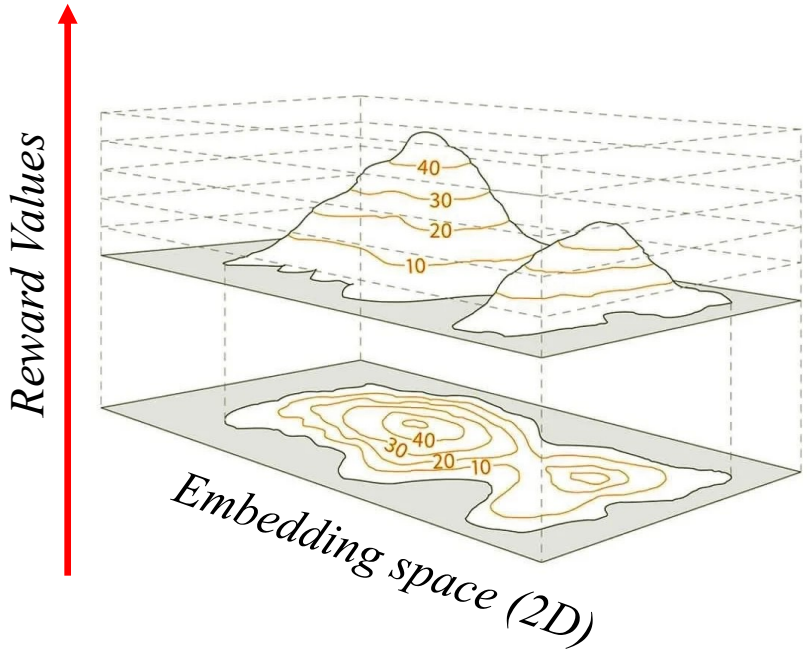
Active Reward Modeling

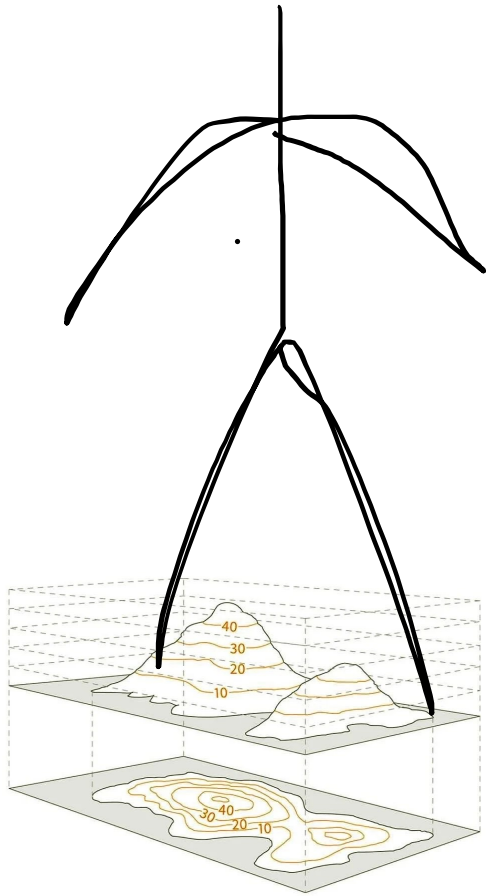
- RM: Constructing elevation maps
  - Preference-based RM: using pairwise comparisons



Active Reward Modeling

- RM: Constructing elevation maps
  - Preference-based RM: using pairwise comparisons





- Theory:
  - Guided by Fisher-Information [PILAF, Feng et al. ICML'2025]

Local + Global comparisons

Learning Reward Models from Data

IRL for Conversational AI

IRL for Math Reasoning

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Elicit Reasoning Abilities…

2205.11916	• Prompt Optimization (Zero-shot CoT)
2309.03409	• RM for Prompt Optimization
2305.20050	• Process Reward Models
2312.08935	• Process Reward Model through MCTS
2501.12948	• DeepSeek-R1
2504.20571	• RL with 1-sample
2506.10947	• RL wrong answer
2504.13837	• Base Model – Amplifying the format?
	• RM to identify Format?

RM from Binary Data: IRL for *Prompt Optimization*

Prompt-OIRL [Sun et al., 2024]

- Prompt engineering is useful, but empirical…

(User Directly Asks the Question)
a^4 = 1, what is a?

(GPT-4 gives the *correct* answer)
Given the equation $a^4 = 1$, we can find the possible values for a. (*…some intermediate steps…*) So, a could be 1, -1, i, or $-i$.

(User Uses Multi Agent Debate Prompting)
a^4 = 1, what is a? Two experts are debating on the answer:

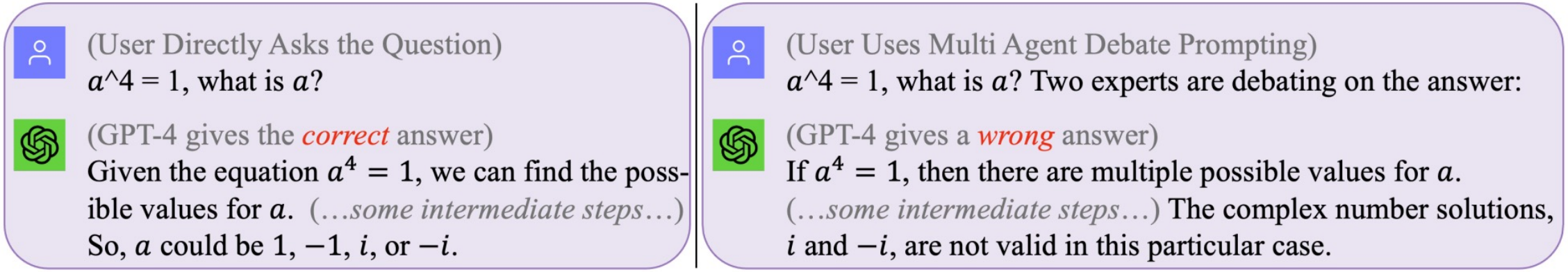
(GPT-4 gives a *wrong* answer)
If $a^4 = 1$, then there are multiple possible values for a. (*…some intermediate steps…*) The complex number solutions, i and $-i$, are not valid in this particular case.

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

Prompt-OIRL [Sun et al., 2024]

- Prompt engineering is useful, but empirical…



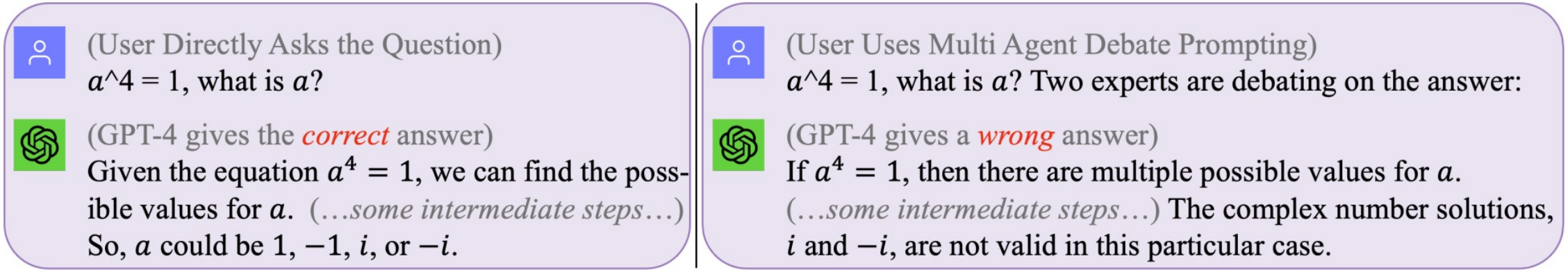
- Automatic prompt engineering using RL?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

Prompt-OIRL [Sun et al., 2024]

- Prompt engineering is useful, but empirical…



- Automatic prompt engineering using RL?
  - Huge vocabulary space
  - Expensive
  - Prompt-Dependent

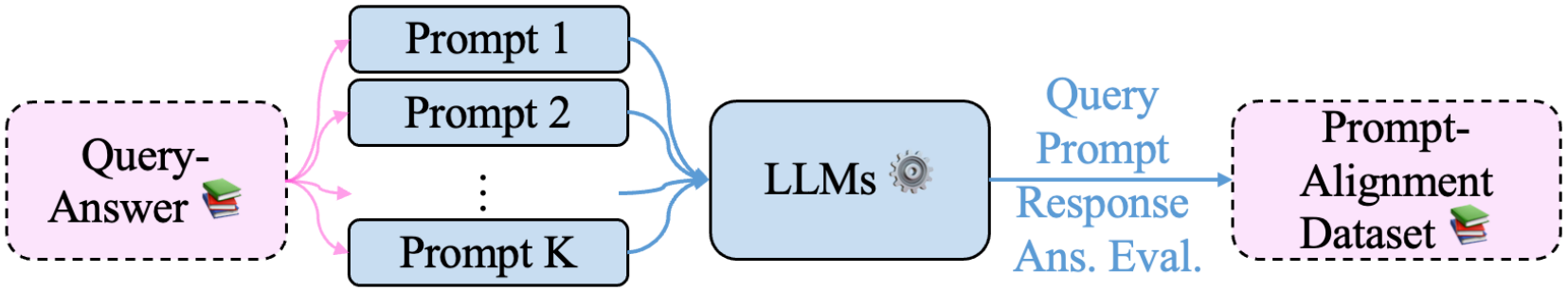
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

Prompt-OIRL [Sun et al., 2024]

- Learning from demonstrative behaviors can be more efficient!

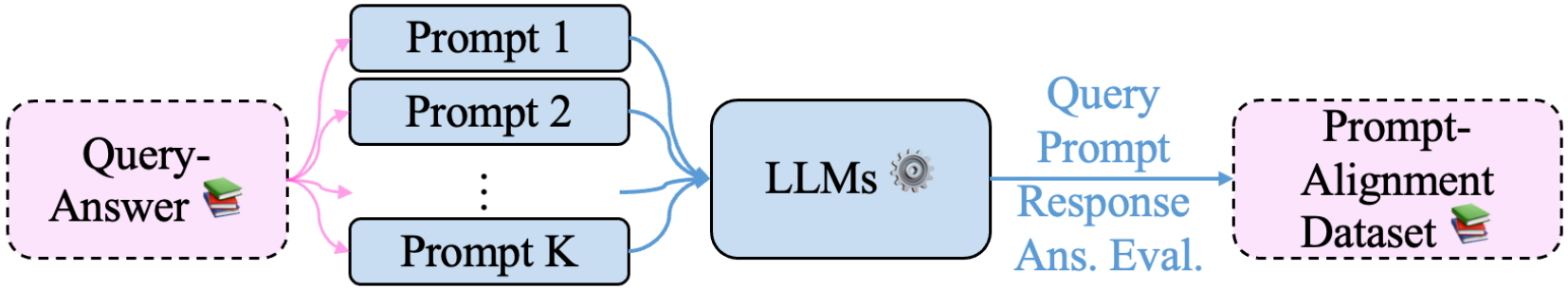
- Prompt Optimization as Inverse RL: learning from expert prompts



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

- Prompt-alignment dataset: [Query, Prompt, Response, Correctness]

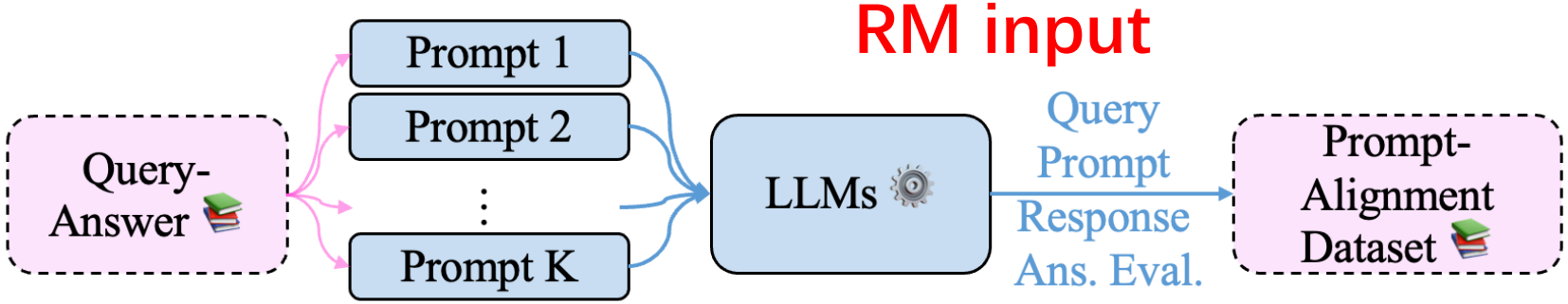


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

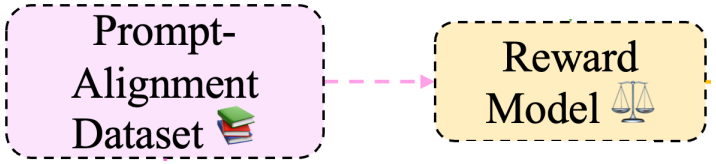
RM from Binary Data: IRL for *Prompt Optimization*

RM input RM output

- Prompt-alignment dataset: [**Query, Prompt**, Response, **Correctness**]



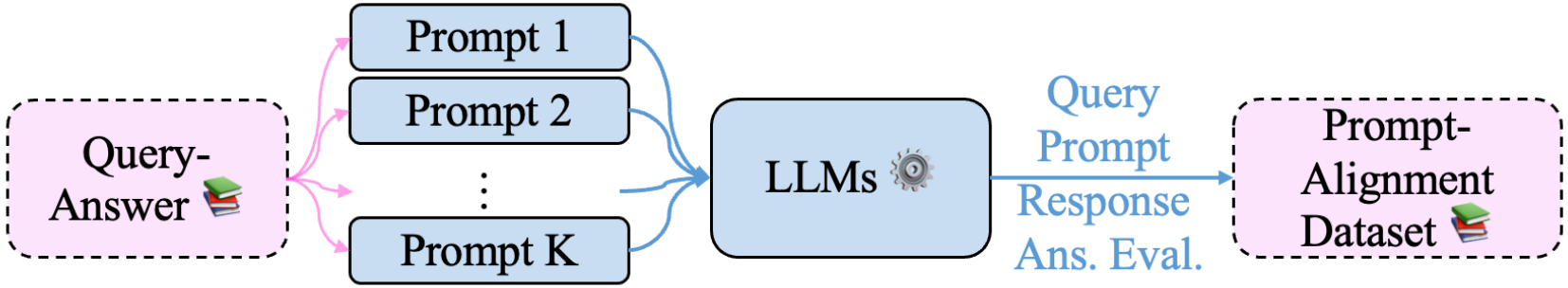
- Inverse RL:
  - (training) Reward Model



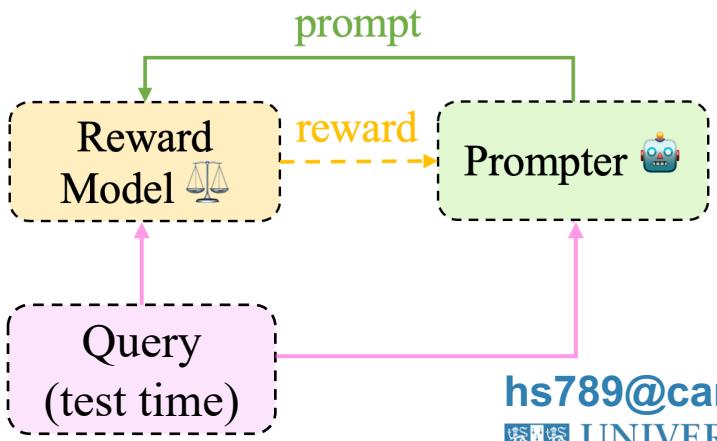
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

- Prompt-alignment dataset: [**Query, Prompt**, Response, **Correctness**]

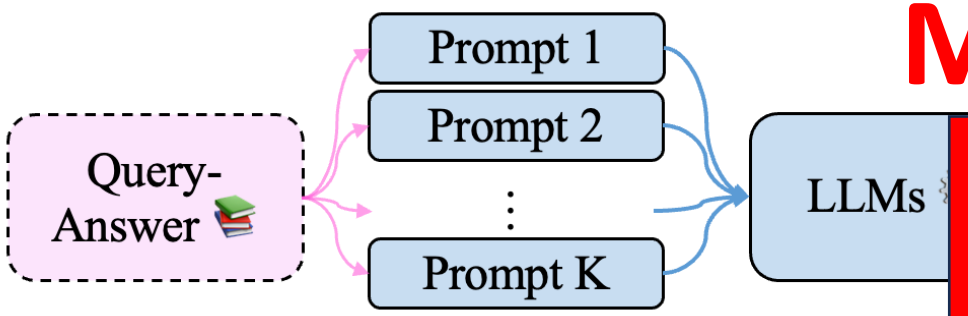


- Inverse RL:
  - (training) Reward Model
  - (test-time) Prompt Optimization
    select the *best* prompt for each query using RM

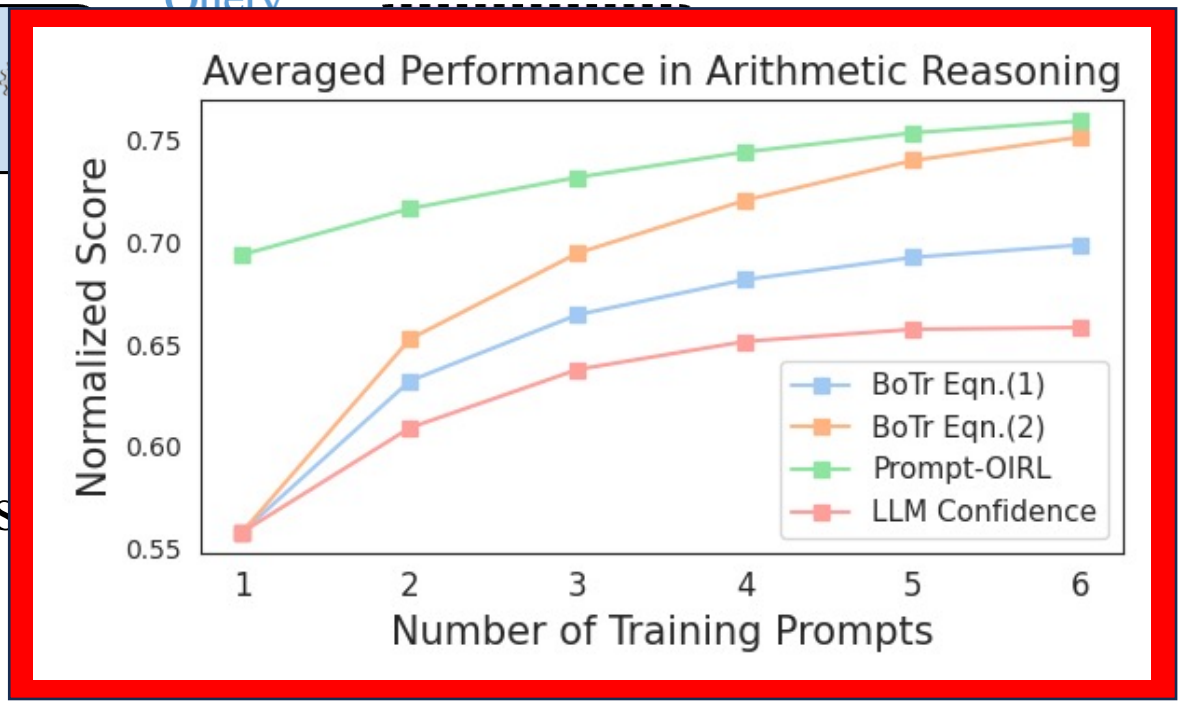


RM from Binary Data: IRL for *Prompt Optimization*

- Prompt-alignment dataset: [**Query, Prompt**, Response, **Correctness**]



Math Reasoning Ability



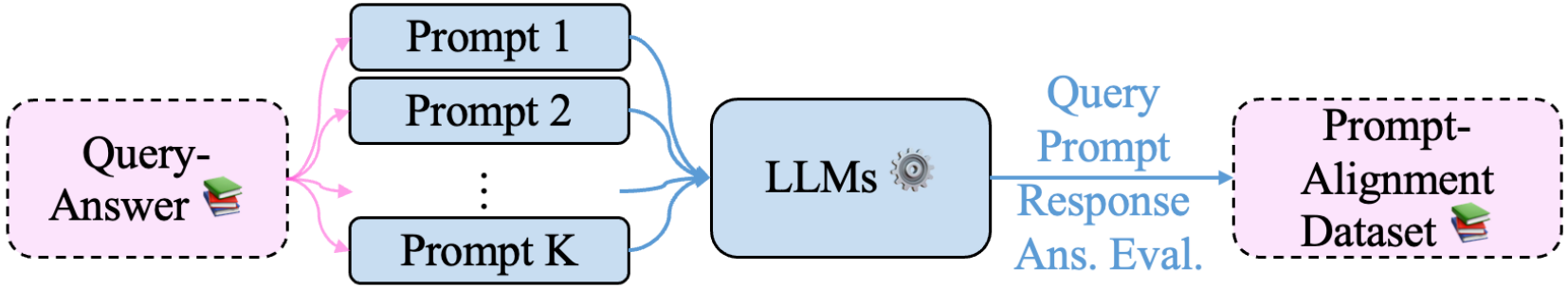
- Inverse RL:
  - (training) Reward Model
  - (test-time) Prompt Optimization
    select the *best* prompt for each query us

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

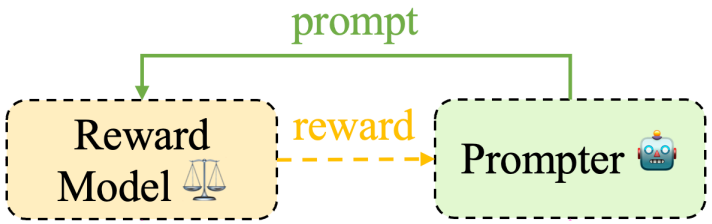
RM --- Transfer to Reasoning Models?

Math Reasoning

- ~~Prompt-alignment~~ dataset: [**Query,** ~~**Prompt**~~**, Thoughts, Correctness**]



- Inverse RL:
  - (training) Reward Model
  - (test-time) Prompt Optimization
    select the *best* response for each query using RM



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Part 4:

LLM Optimization with Reward Models

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Optimization Algorithms

- MDP\R + $\widehat{R}$ = MDP
- Any RL algorithm can be applied

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Optimization Algorithms

- MDP\R + $\hat{R}$ = MDP
- Any RL algorithm can be applied
  - Best-of-N (Academic Research)

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

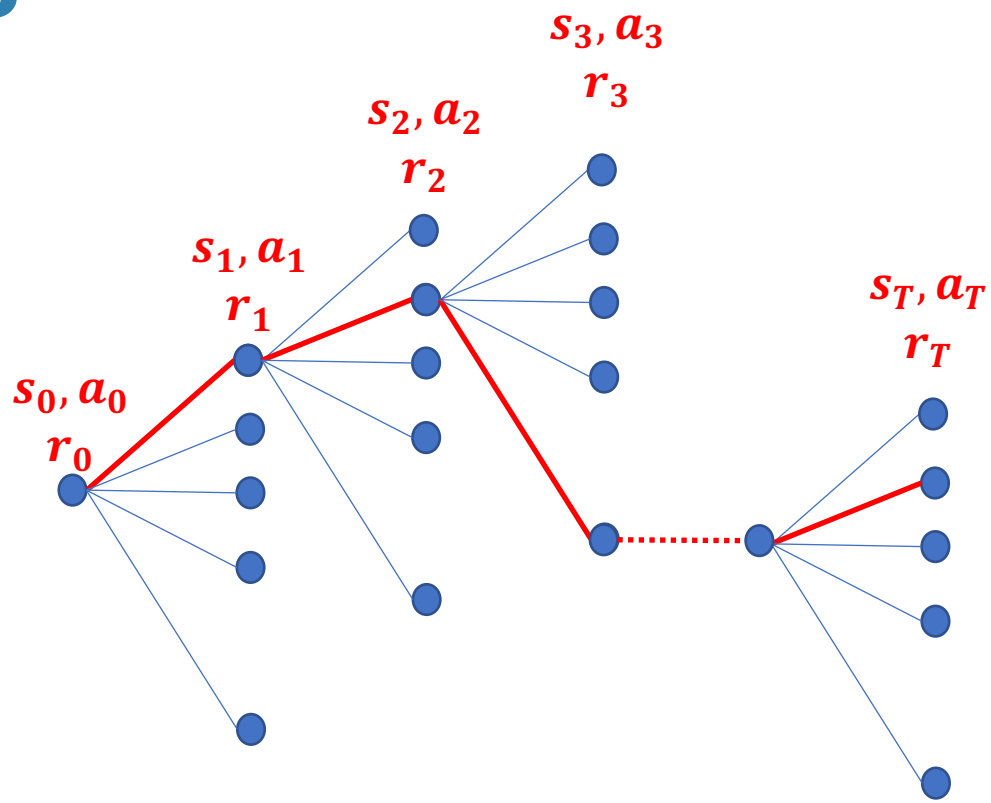
Optimization Algorithms

- MDP\R + $\widehat{R}$ = MDP
- Any RL algorithm can be applied
  - Best-of-N (Academic Research)
  - Iterative Rejection-Sampling

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

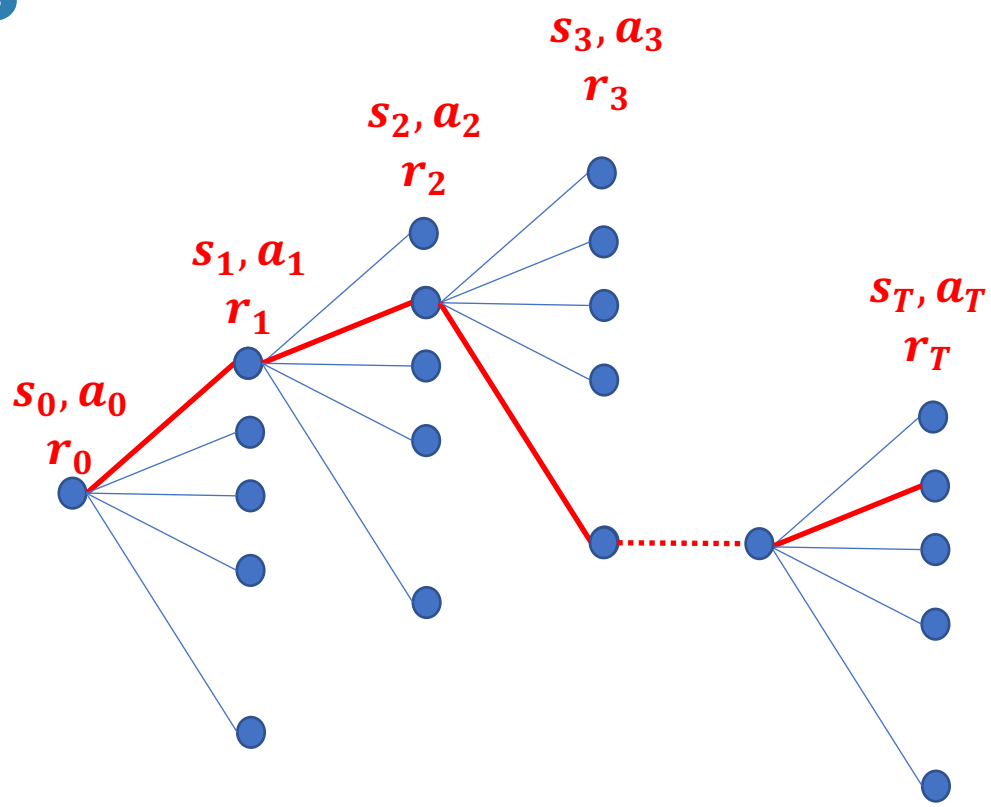
Optimization Algorithms

- MDP\R + $\widehat{R}$ = MDP
- Any RL algorithm can be applied
  - Best-of-N (Academic Research)
  - Iterative Rejection-Sampling
  - PPO (GAE: MC+TD)
  - REINFORCE/GRPO (Monte-Carlo)



Optimization Algorithms

- MDP\R + $\widehat{\mathrm{R}}$ = MDP
- Any RL algorithm can be applied
  - Best-of-N (Academic Research)
  - Iterative Rejection-Sampling
  - PPO (GAE: MC+TD)
  - REINFORCE/GRPO (Monte-Carlo)



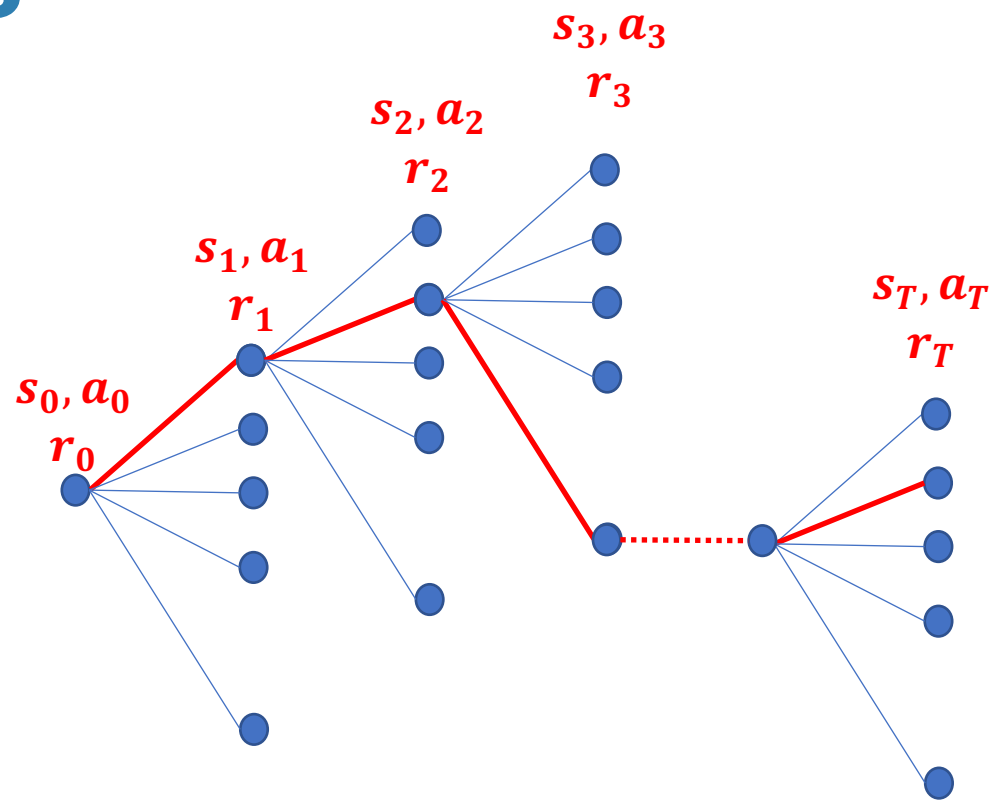
MC Value Estimation: $\mathrm{Q}(s_0, a_0) \leftarrow r_0 + \mathbb{E}_{a_t}[\sum_{t=1}^{T} r_t]$

TD Value Estimation: $\mathrm{Q}(s_0, a_0) \leftarrow r_0 + \mathbb{E}_{a'}[Q(s_1, a')]$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Optimization Algorithms

- MDP\R + $\widehat{R}$ = MDP
- Any RL algorithm can be applied
  - Best-of-N (Academic Research)
  - Iterative Rejection-Sampling
  - PPO (GAE: MC+TD)
  - REINFORCE/GRPO (Monte-Carlo)
    [ReMax, Li et al., 2023]



MC Value Estimation: $\mathrm{Q}(s_0, a_0) \leftarrow r_0 + \mathbb{E}_{a_t}[\sum_{t=1}^{T} r_t] = r_T$ (traj.-level reward)

TD Value Estimation: $\mathrm{Q}(s_0, a_0) \leftarrow r_0 + \mathbb{E}_{a'}[Q(s_1, a')] = \cdots = \mathbb{E}_{\tau}[Q(\tau, a = \mathrm{EOS})]$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Optimization Algorithms

- MDP\R + $\widehat{R}$ = MDP
- Any RL algorithm can be applied
  - Best-of-N (Academic Research)
  - Iterative Rejection-Sampling
  - PPO (GAE: MC+TD)
  - REINFORCE/GRPO (Monte-Carlo)
- Reward-guided Decoding

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Optimization Algorithms

- MDP\R + $\hat{R}$ = MDP
- Any RL algorithm can be applied
  - Best-of-N (Academic Research)
  - Iterative Rejection-Sampling
  - PPO (GAE: MC+TD)
  - REINFORCE/GRPO (Monte-Carlo)
- Reward-guided Decoding
- Direct Preference Optimization

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

DPO: Implicit Reward Models

- RLHF as Inverse RL:
  - RM learning objective:

$$\text{loss}(r_\theta) = E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))] \quad \text{(Eqn1)}$$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

DPO: Implicit Reward Models

- RLHF as Inverse RL:
  - RM learning objective:

$$\text{loss}(r_\theta) = E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))] \quad \text{(Eqn1)}$$

  - Policy learning objective:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(y \mid x) \,||\, \pi_{\text{ref}}(y \mid x)\right],$$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

DPO: Implicit Reward Models

- RLHF as Inverse RL:
  - RM learning objective:

$$\text{loss}(r_\theta) = E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))]$$

  - Policy learning objective:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(y \mid x) \,||\, \pi_{\text{ref}}(y \mid x)\right],$$

  - DPO: Closed-form solution for the ***Policy learning objective*** exists!

$$\pi_r(y \mid x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x,y)\right),$$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

DPO: Implicit Reward Models

- RLHF as Inverse RL:
  - RM learning objective:

$$\text{loss}(r_\theta) = E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))]$$

  - Policy learning objective:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(y \mid x) \,||\, \pi_{\text{ref}}(y \mid x)\right],$$

  - DPO:

$$\pi_r(y \mid x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x,y)\right), \qquad r(x,y) = \beta\log\frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta\log Z(x). \quad \textbf{(Eqn2)}$$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

DPO: Implicit Reward Models

- RLHF as Inverse RL:
  - RM learning objective:

$$\text{loss}(r_\theta) = E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))] \quad \text{(Eqn1)}$$

  - Expressing reward with policy:

$$r(x,y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x). \quad \text{(Eqn2)}$$

  - DPO: *your language model is secretly a reward model*

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

DPO: Implicit Reward Models

- RLHF as Inverse RL:
  - RM learning objective:

$$\text{loss}(r_\theta) = E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))] \quad \text{(Eqn1)}$$

  - Expressing reward with policy:

$$r(x,y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x). \quad \text{(Eqn2)}$$

  - DPO: *your language model is secretly a reward model*

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

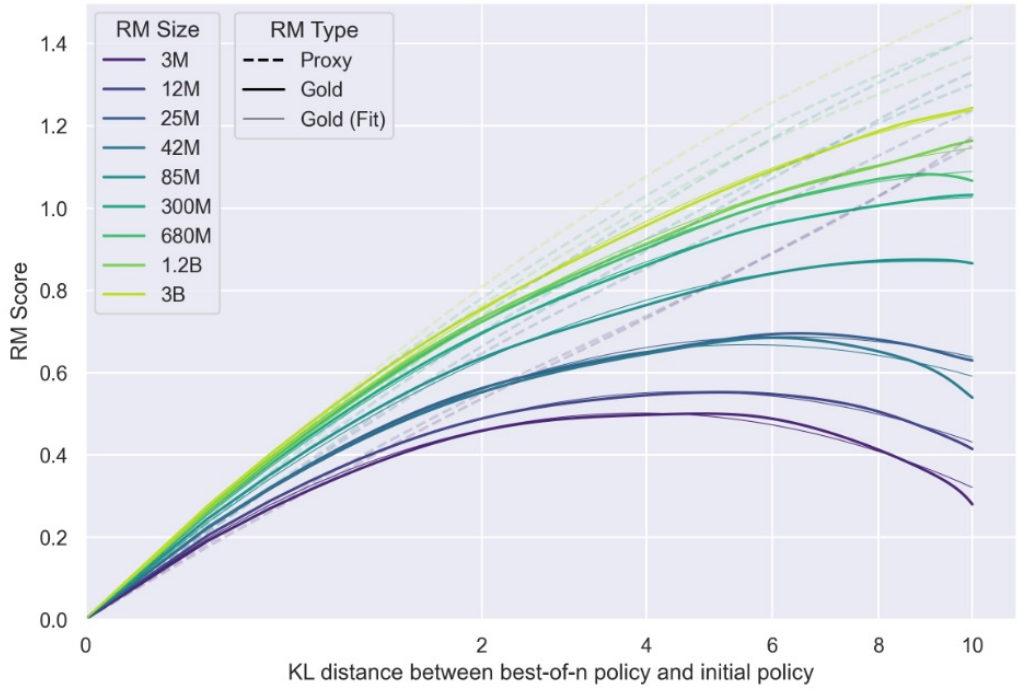
Comparing DPO and PPO

- Performance-wise: PPO > DPO [Xu et al., 2024]
- Easy-to-Implement: DPO > PPO
- Trade-off: Iterative DPO / Online DPO [Xiong et al., 2023]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Challenges and Opportunities

- Reward Model overoptimization [Gao et al., 2022]



(a) BoN

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

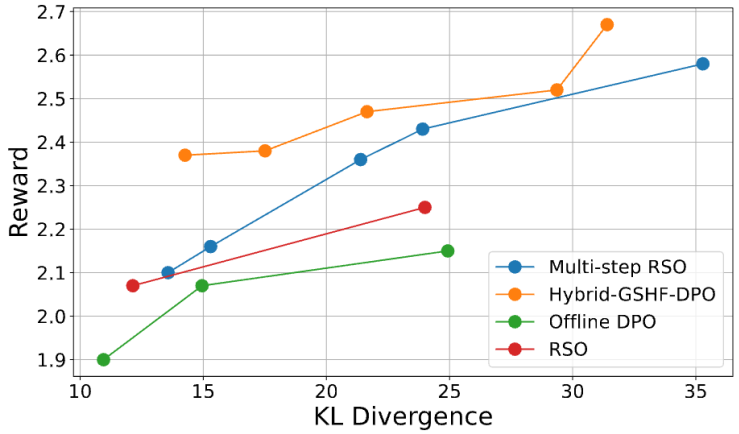
Challenges and Opportunities

- Reward Model overoptimization [Gao et al., 2022]
  - RM ensemble [Coste et al., 2024] [Ahmed et al., 2024] [Zhang et al., 2024]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

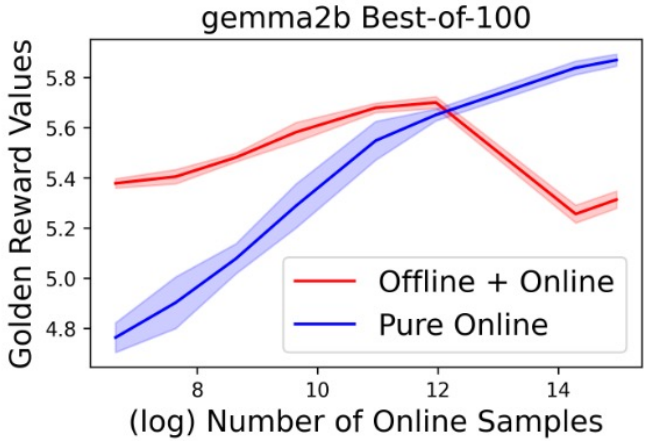
Challenges and Opportunities

- Reward Model overoptimization [Gao et al., 2022]
  - RM ensemble [Coste et al., 2024] [Ahmed et al., 2024] [Zhang et al., 2024]
- On-policy/Off-policy annotations
  - Iterative online annotation improves efficiency [Xiong et al. 2024]



Challenges and Opportunities

- Reward Model overoptimization [Gao et al., 2022]
  - RM ensemble [Coste et al., 2024] [Ahmed et al., 2024] [Zhang et al., 2024]
- On-policy/Off-policy annotations
  - Iterative online annotation improves efficiency [Xiong et al. 2024]
  - Less can be more with online preference [Sun et al., 2024] [Zhou et al., 2023]



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Challenges and Opportunities

- Reward Model overoptimization [Gao et al., 2022]
  - RM ensemble [Coste et al., 2024] [Ahmed et al., 2024] [Zhang et al., 2024]
- On-policy/Off-policy annotations
  - Iterative online annotation improves efficiency [Xiong et al. 2024]
  - Less can be more with online preference [Sun et al., 2024] [Zhou et al., 2023]
- How to effectively collect preference data?
  - Max-Margin/Entropy [Muldrew et al., 2024]
  - Fisher Information [Feng et al., 2025] [Shen et al., 2025]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Challenges and Opportunities

- Reward Model overoptimization [Gao et al., 2022]
  - RM ensemble [Coste et al., 2024] [Ahmed et al., 2024] [Zhang et al., 2024]
- On-policy/Off-policy annotations
  - Iterative online annotation improves efficiency [Xiong et al. 2024]
  - Less can be more with online preference [Sun et al., 2024] [Zhou et al., 2023]
- How to effectively collect preference data?
  - Max-Margin/Entropy [Muldrew et al., 2024]
  - Fisher Information [Feng et al., 2025] [Shen et al., 2025]
- Other data type?
  - Critique data [Zhang et al., 2025] [Ankner et al., 2024] [Wu et al., 2024]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Challenges and Opportunities

The Challenge of *Generalization*

- Reward Model overoptimization [Gao et al., 2022]
  - RM ensemble [Coste et al., 2024] [Ahmed et al., 2024] [Zhang et al., 2024]
- On-policy/Off-policy annotations
  - Iterative online annotation improves efficiency [Xiong et al. 2024]
  - Less can be more with online preference [Sun et al., 2024] [Zhou et al., 2023]
- How to effectively collect preference data?
  - Max-Margin/Entropy [Muldrew et al., 2024]
  - Fisher Information [Feng et al., 2025] [Shen et al., 2025]
- Other data type?
  - Critique data [Zhang et al., 2025] [Ankner et al., 2024] [Wu et al., 2024]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Part 5:

Insights from the Sparse-Reward RL Literature

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

What are Unique with LLMs?

Task	Action Space	State Space	Reward Signal	Method	Transition
Atari	Disc. ~ 1e1	Image	Dense (mostly)	DQN	Unknown
Atari-Explore	Disc. ~ 1e1	Image	Sparse	Curiosity-Driven	Unknown
Go	Disc. ~ 1e2	Disc. ~1e100	Sparse	MCTS, Self-Play	Known
Dota2	Disc. ~ 1e6	Partial Observable	Dense & Sparse	(MA)PPO	Unknown
StarCraft	Disc. ~ 1e26	Partial Observable	Dense & Sparse	BC, AC, League	Unknown
Multi-Goal	Cont. Dim~1e2	Cont. Dim~1e2	Sparse	Hindsight Exp. Replay	Unknown
Reasoning	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Sparse**	GRPO	**Known**
Alignment	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Noisy Preference**	PPO, DPO, REINFORCE	**Known**

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Transferrable Insights?

Task	Action Space	State Space	Reward Signal	Method	Transition
Atari	Disc. ~ 1e1	Image	Dense (mostly)	DQN	Unknown
Atari-Explore	Disc. ~ 1e1	Image	Sparse	Curiosity-Driven	Unknown
Go	Disc. ~ 1e2	Disc. ~1e100	Sparse	MCTS, Self-Play	Known
Dota2	Disc. ~ 1e6	Partial Observable	Dense & Sparse	(MA)PPO	Unknown
StarCraft	Disc. ~ 1e26	Partial Observable	Dense & Sparse	BC, AC, League	Unknown
Multi-Goal	Cont. Dim~1e2	Cont. Dim~1e2	Sparse	Hindsight Exp. Replay	Unknown
Reasoning	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Sparse**	GRPO	**Known**
Alignment	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Noisy Preference**	PPO, DPO, REINFORCE	**Known**

Learning from Failure?

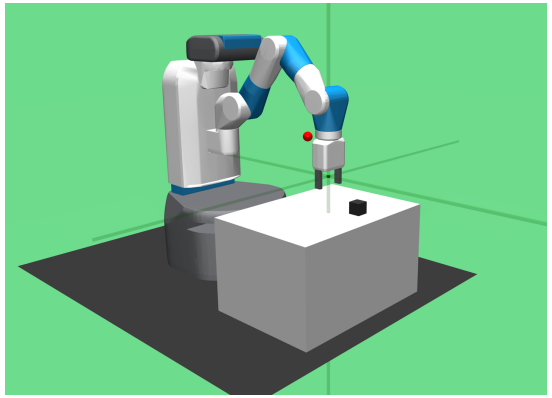
Reward Shaping?

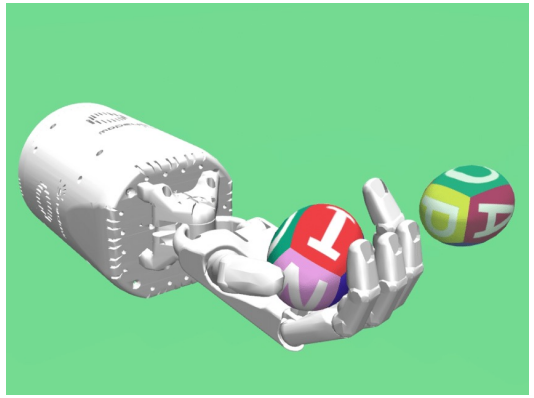
Self-Play?

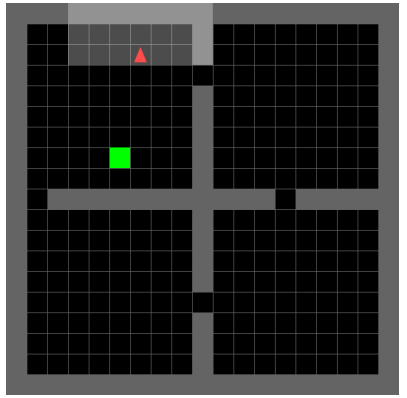
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- Multi-Goal task
- Failing in achieve a certain goal = successfully achieving another goal







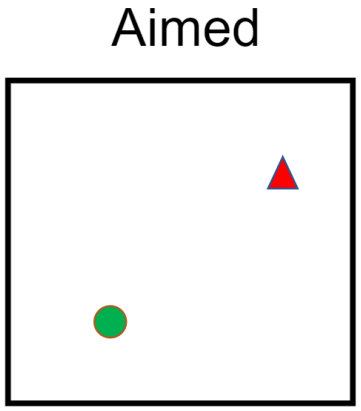
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

learn how to reach goal g : $\pi(\cdot\,|s_0, g)$

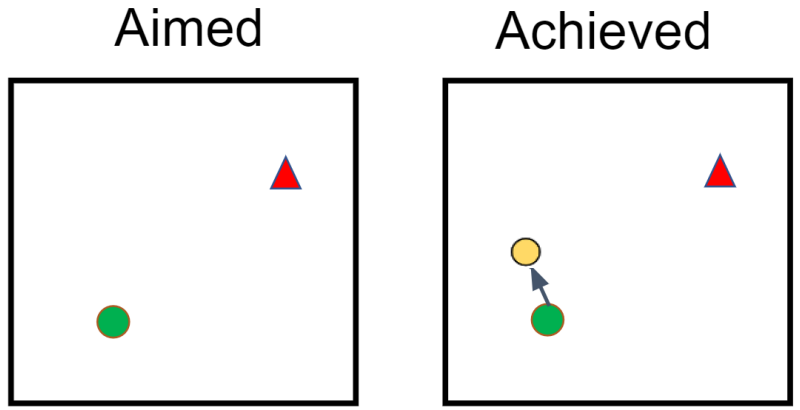


Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

learn how to reach goal g : $\pi(\cdot\,|s_0, g)$



$a_1,$

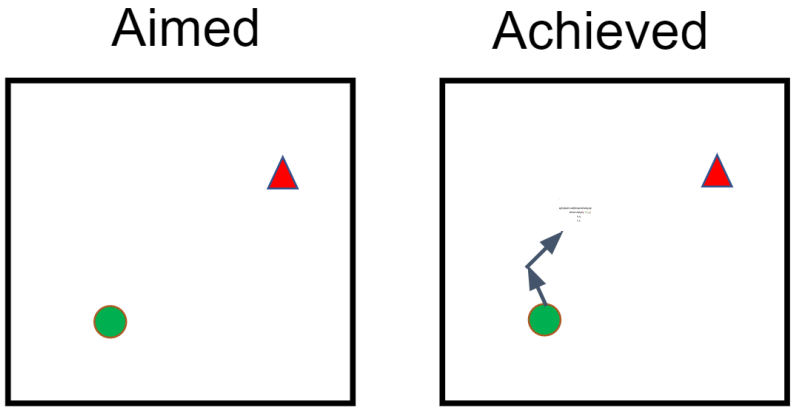
$s_1,$

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

learn how to reach goal g : $\pi(\cdot\,|s_0, g)$



$a_1, a_2,$

$s_1, s_2,$

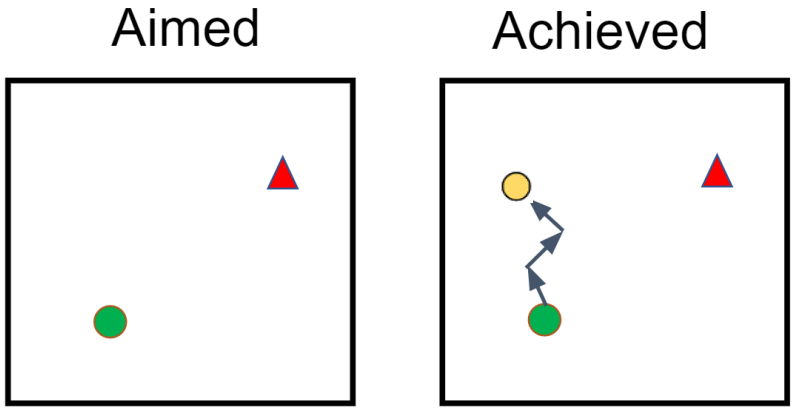
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

learn how to reach goal g : $\pi(\cdot\,|s_0, g)$



a_1, a_2, a_3

s_1, s_2, s_3

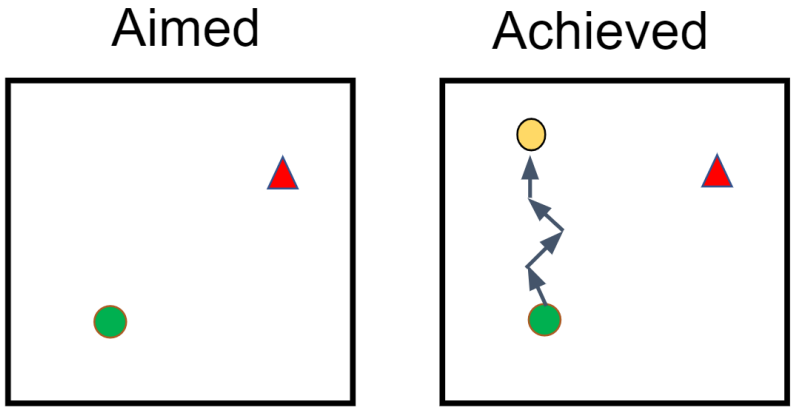
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

learn how to reach goal g : $\pi(\cdot\,|s_0, g)$



a_1, a_2, a_3, a_4

s_1, s_2, s_3, s_4

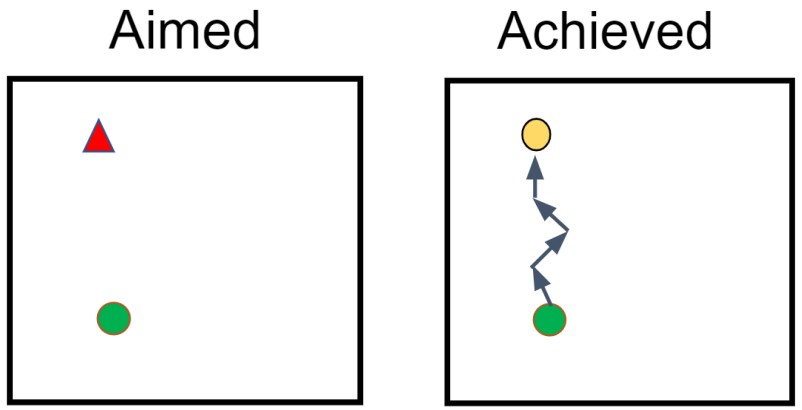
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

learn how to reach goal g : $\pi(\cdot\,|s_0, g)$



a_1, a_2, a_3, a_4

s_1, s_2, s_3, s_4

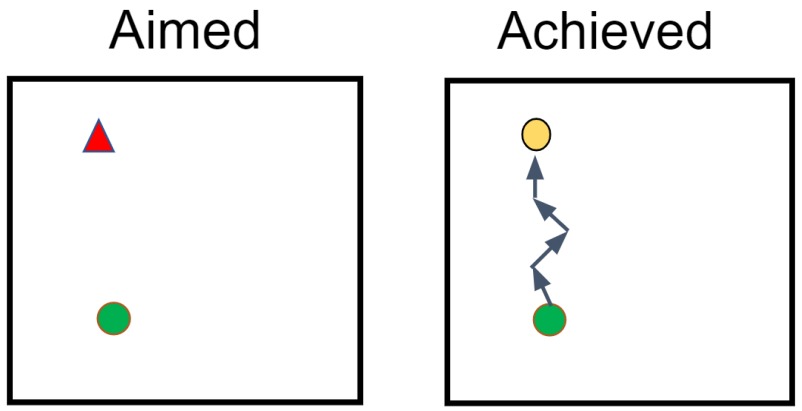
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

learn how to reach goal g : $\pi(\cdot\,|s_0, g)$



a_1, a_2, a_3, a_4

s_1, s_2, s_3, s_4

$a_1 \leftarrow \pi(\cdot\,|s_0, g)$

$a_2 \leftarrow \pi(\cdot\,|s_1, g)$

$a_3 \leftarrow \pi(\cdot\,|s_2, g)$

$a_4 \leftarrow \pi(\cdot\,|s_3, g)$

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:
  Hindsight Self-Imitate Learning [Sun et al., 2019]

- **Multi-goal** tasks in LLM alignment?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

  Hindsight Self-Imitate Learning [Sun et al., 2019]

- **Multi-goal** tasks in LLM alignment?
  LLM Web Agent [He et al., 2024]
  more to explore!

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Reward Shaping (and Credit Assignment)

- Given *delayed* reward, assign credit to each action?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Reward Shaping (and Credit Assignment)

- Given *delayed* reward, assign credit to each action?
  - Credit Assignment Problem [Sutton, 1984] Recent Survey [Pignatelli et al., 2024]
  - Return Decomposition [Arjona-Medina et al., 2019][Ren et al., 2022]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Reward Shaping (and Credit Assignment)

- Given *delayed* reward, assign credit to each action?
  - Credit Assignment Problem [Sutton, 1984] Recent Survey [Pignatelli et al., 2024]
  - Return Decomposition [Arjona-Medina et al., 2019][Ren et al., 2022]
- In LLM alignment:
  - Free-of-anno.: Token-level dense reward using attention [Chan et al., 2024]
  - Free-of-anno.: PRMs from Monte-Carlo Tree Search [Wang et al. 2023]
  - Anno.: Process-supervised Reward Models (PRMs) [Lightman et al., 2023]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Reward Shaping (and Credit Assignment)

- Given *delayed* reward, assign credit to each action?
  - Credit Assignment Problem [Sutton, 1984] Recent Survey [Pignatelli et al., 2024]
  - Return Decomposition [Arjona-Medina et al., 2019][Ren et al., 2022]
- In LLM alignment:
  - Free-of-anno.: Token-level dense reward using attention [Chan et al., 2024]
  - Free-of-anno.: PRMs from Monte-Carlo Tree Search [Wang et al. 2023]
  - Anno.: Process-supervised Reward Models (PRMs) [Lightman et al., 2023]
- Conservative reward shaping keeps the optimal policy [Ng et al., 1999]
- Reward shaping can change learning behavior [Sun et al., 2022]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Best Reward: Dense or Sparse?

- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Best Reward: Dense or Sparse?

- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better? --- Not always

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

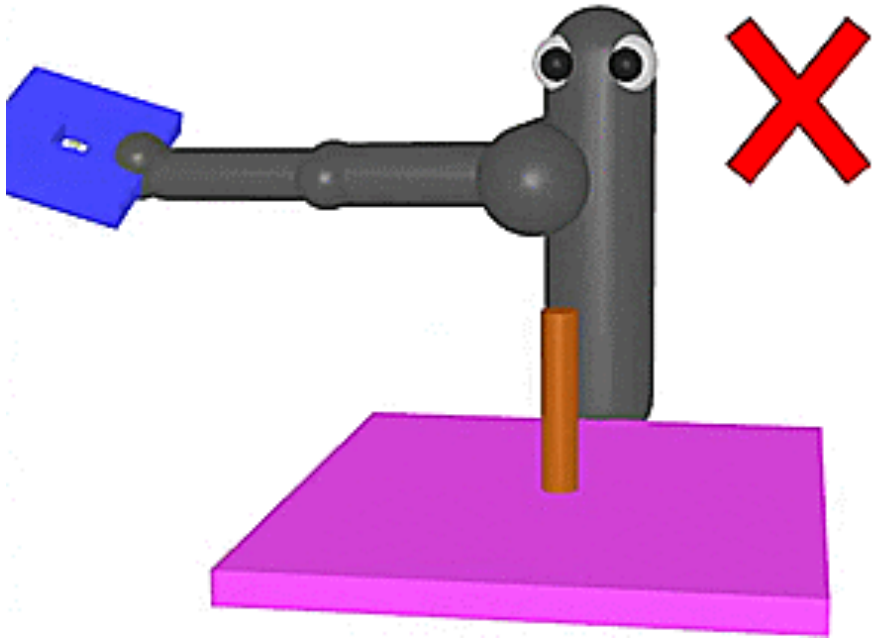
Best Reward: Dense or Sparse?

- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better? --- Not always
  - 1. DeepSeek-R1 [DeepSeek-AI, 2025]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Best Reward: Dense or Sparse?

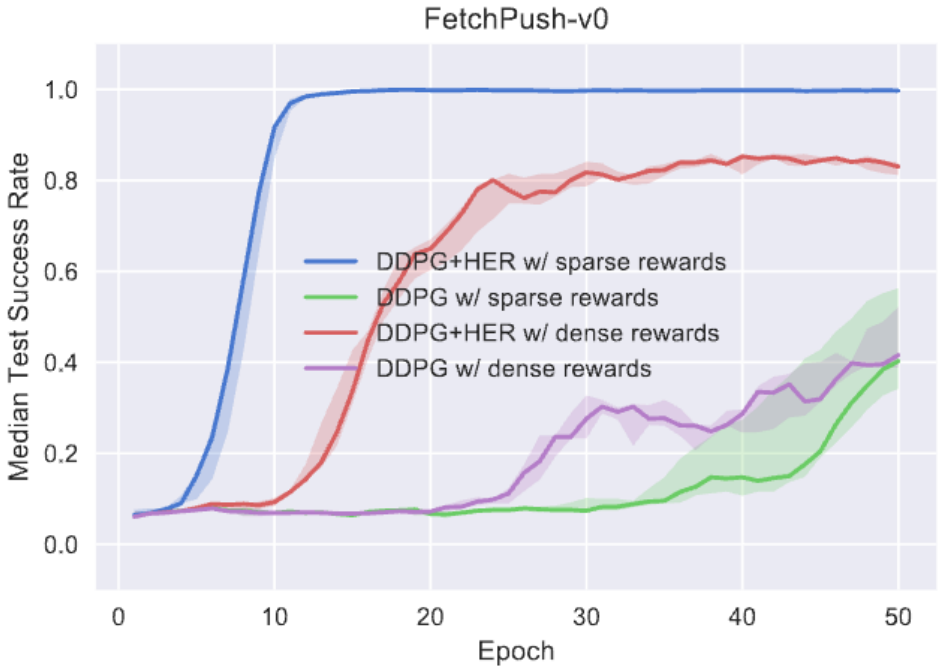
- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better? --- Not always
  - 1. DeepSeek-R1 [DeepSeek-AI, 2025]
  - 2. Suboptimality [Florensa et al., 2017]



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Best Reward: Dense or Sparse?

- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better? --- Not always
  - 1. DeepSeek-R1 [DeepSeek-AI, 2025]
  - 2. Suboptimality [Florensa et al., 2017]
  - 3. In multi-goal tasks [Plappert et al., 2018]



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Best Reward: Dense or Sparse?

- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better? --- Not always
  - 1. DeepSeek-R1 [DeepSeek-AI, 2025]
  - 2. Suboptimality [Florensa et al., 2017]
  - 3. In multi-goal tasks [Plappert et al., 2018]
- Do we need dense reward?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Best Reward: Dense or Sparse?

- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better? --- Not always
  - 1. DeepSeek-R1 [DeepSeek-AI, 2025]
  - 2. Suboptimality [Florensa et al., 2017]
  - 3. In multi-goal tasks [Plappert et al., 2018]
- Do we need dense reward?
  - Warm-start from demo [Nair et al., 2017]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Best Reward: Dense or Sparse?

- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better? --- Not always
  - 1. DeepSeek-R1 [DeepSeek-AI, 2025]
  - 2. Suboptimality [Florensa et al., 2017]
  - 3. In multi-goal tasks [Plappert et al., 2018]
- Do we need dense reward?
  - No --- Warm-start from demo [Nair et al., 2017]
  - Yes --- AlphaGo

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Self-Play

- What can we learn from the success of AlphaGo (Zero)?

Self-Play

- What can we learn from the success of AlphaGo (Zero)?
- Why AlphaGo?

Task	Action Space	State Space	Reward Signal	Method	Transition
Go	Disc. ~ 1e2	Disc. ~1e100	Sparse	MCTS, Self-Play	Known
Reasoning	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Sparse**	GRPO	**Known**
Alignment	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Noisy Preference**	PPO, DPO, REINFORCE	**Known**

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Self-Play

- What can we learn from the success of AlphaGo (Zero)?
- Why AlphaGo?

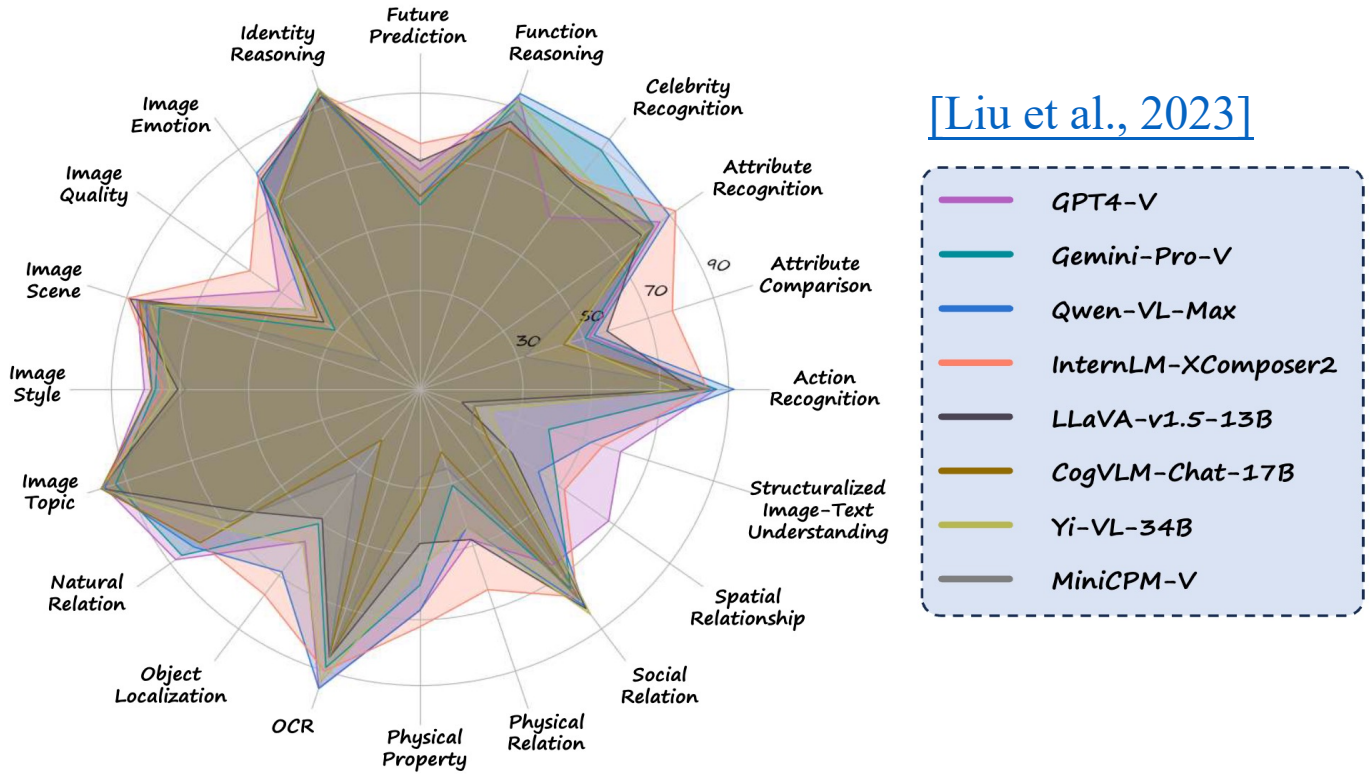
Task	Action Space	State Space	Reward Signal	Method	Transition
Go	Disc. ~ 1e2	Disc. ~1e100	Sparse	MCTS, Self-Play	Known
Reasoning	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Sparse**	GRPO	**Known**
Alignment	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Noisy Preference**	PPO, DPO, REINFORCE	**Known**

- Self-Play are also widely used in LLM alignment / reasoning

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Self-Play

- Self-Play in LLMs: self-critique/-correction/-evaluation
  - LLMs are good at *some* aspects than *others*
  - Improvement is not guaranteed / bounded



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Self-Play

- Self-Play in LLMs: self-critique/-correction/-evaluation
  - LLMs are good at *some* aspects than *others*
  - Improvement is not guaranteed / bounded

- Self-Play in Go(Shogi/Chess)/StarCraft:
  - (nearly-symmetric) zero-sum two player games
  - Winning the old policy is improvement (by definition)

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Self-Play

- Self-Play in LLMs: self-critique/-correction/-evaluation
  - LLMs are good at *some* aspects than *others*
  - Improvement is not guaranteed / bounded

- Self-Play in Go(Shogi/Chess)/StarCraft:
  - (nearly-symmetric) zero-sum two player games
  - Winning the old policy is improvement (by definition)

- Real Self-Play in games?
  - [Cheng et al., 2024] [Hu et al., 2024] [Duan et al., 2025]

Takeaways

- In Multi-Goal tasks, Hindsight methods have great potential
- 3 methods for dense reward:
  - Free using attention weights (stability)
  - Free using MCTS (not as good as using sparse reward)
  - Extra annotation (?)
- Dense reward not designed properly may lead to suboptimal
- In AlphaGo, searching is needed to achieve super-human performance
- Self-Play can be powerful

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Takeaways

- In Multi-Goal tasks, Hindsight methods have great potential
- 3 methods for dense reward:
  - Free using attention weights (stability)
  - Free using MCTS (not as good as using sparse reward)
  - Extra annotation (?)
- Dense reward not designed properly may lead to suboptimal
- In AlphaGo, searching is needed to achieve super-human performance
- Self-Play can be powerful

Future of IRL x LLMs

The AlphaGo moment of LLMs will come ☺

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Thank you!

sites.google.com/view/irl-llm

hs789@cam.ac.uk