Inverse Reinforcement Learning Meets LLM Alignment

AAAI 2025 Tutorial

Mihaela van der Schaar, Hao Sun

Feb. 2025

van_der_Schaar
\ LAB

vanderschaar-lab.com

hs789@cam.ac.uk

@HolarisSun

sites.google.com/view/irl-llm

Content

- Part 0: Motivations
- Part 1: Reinforcement Learning: Forward and Inverse
- Part 2: Learning Reward Models from Data
- Part 3: Infra for IRL-LLM Research
- Part 4: Lessons from Sparse Reward RL

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 0: Motivations
  - 0.1 How to extend the success of DeepSeek-R1
- Part 1: Reinforcement Learning: Forward and Inverse
- Part 2: Learning Reward Models from Data
- Part 3: Infra for IRL-LLM Research
- Part 4: Lessons from Sparse Reward RL

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 0: Motivations
- Part 1: Reinforcement Learning: Forward and Inverse
  - 1.1 RL, MDP; Inverse RL, MDP\R
  - 1.2 LLM alignment as Inverse RL
  - 1.3 Why do we (always) need RMs?
- Part 2: Learning Reward Models from Data
- Part 3: Infra for IRL-LLM Research
- Part 4: Lessons from Sparse Reward RL

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 0: Motivations
- Part 1: Reinforcement Learning: Forward and Inverse
- Part 2: Learning Reward Models from Data
  - 2.1 Reward Modeling in the Wild
  - 2.2 Challenges and Opportunities
- Part 3: Infra for IRL-LLM Research
- Part 4: Lessons from Sparse Reward RL

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 0: Motivations
- Part 1: Reinforcement Learning: Forward and Inverse
- Part 2: Learning Reward Models from Data
- Part 3: Infra for IRL-LLM Research
  - 3.1 Design Principles: Reproducible, Controllable, Scalable
  - 3.2 Efficiency v.s. Performance
  - 3.3 Demonstrative Use Cases
- Part 4: Lessons from Sparse Reward RL

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Content

- Part 0: Motivations
- Part 1: Reinforcement Learning: Forward and Inverse
- Part 2: Learning Reward Models from Data
- Part 3: Infra for IRL-LLM Research
- Part 4: Lessons from Sparse Reward RL
  - 4.1 Hindsight Knowledge
  - 4.2 Reward Shaping, Credit Assignment
  - 4.3 Tradeoffs between Dense Reward (PRM) and Sparse Reward (ORM)
  - 4.4 Power from Self-Play

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Part 0:

Motivations

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Success of Data-Driven Large Models

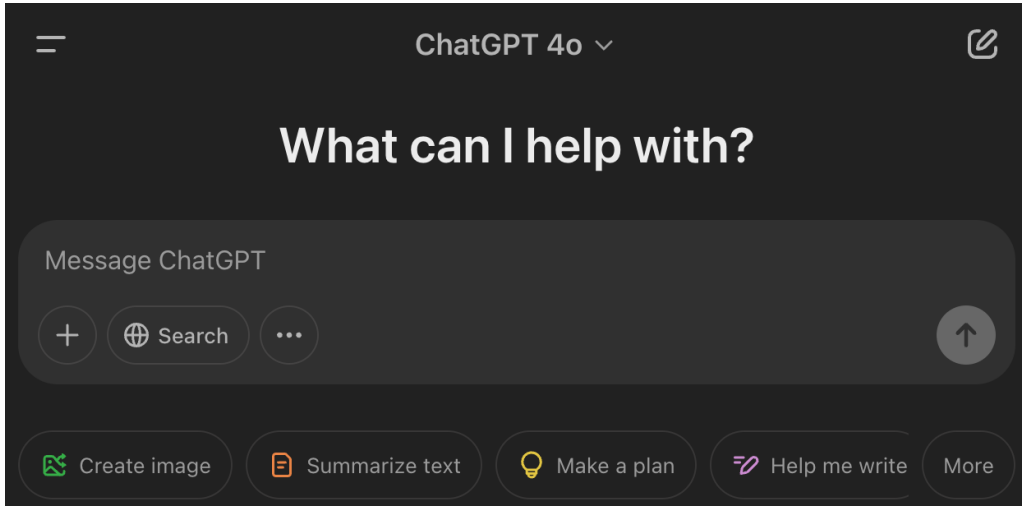
- *Realistic* contents generated by data-driven large models

[Zhang et al., 2025]

Success of Data-Driven Large Models

- Video
- Large Language Models

Video source: openai.com/sora/



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Success of Large-Scale RL

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

Reward ⬆️⬆️

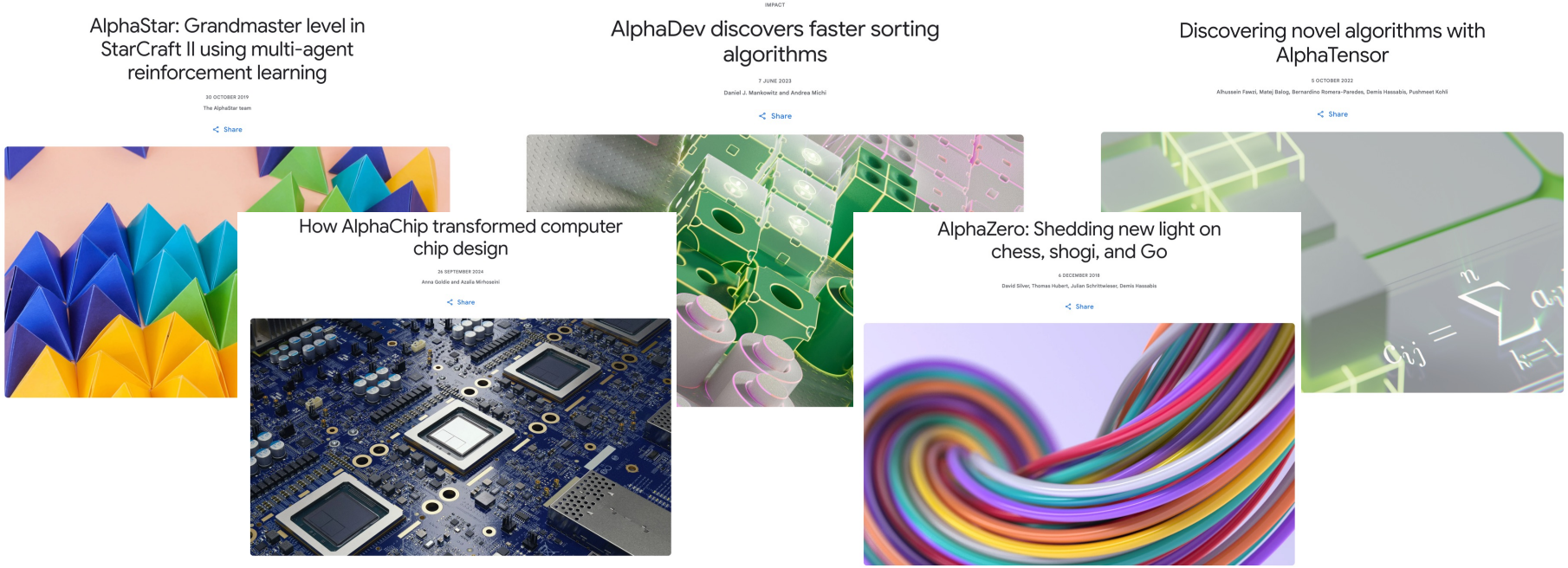


Image source: deepmind.google/discover

openai.com/index/openai-five/

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Success 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

Success from Both Sides

Language (generative) Models

- Understanding and generating
- Fast adaptation to new tasks

Impressive because those contents look like human might generate --- [2](rephrased)

RL

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

Impressive because no person had thought of it --- [2]

[1] Talk: David Silver - Towards Superhuman Intelligence - RLC 2024
[2] Talk: Sergey Levine - Reinforcement Learning in the Age of Foundation Models - 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 + RL:
  More interpretable, assist, empower and inspire human

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Combining the Success?

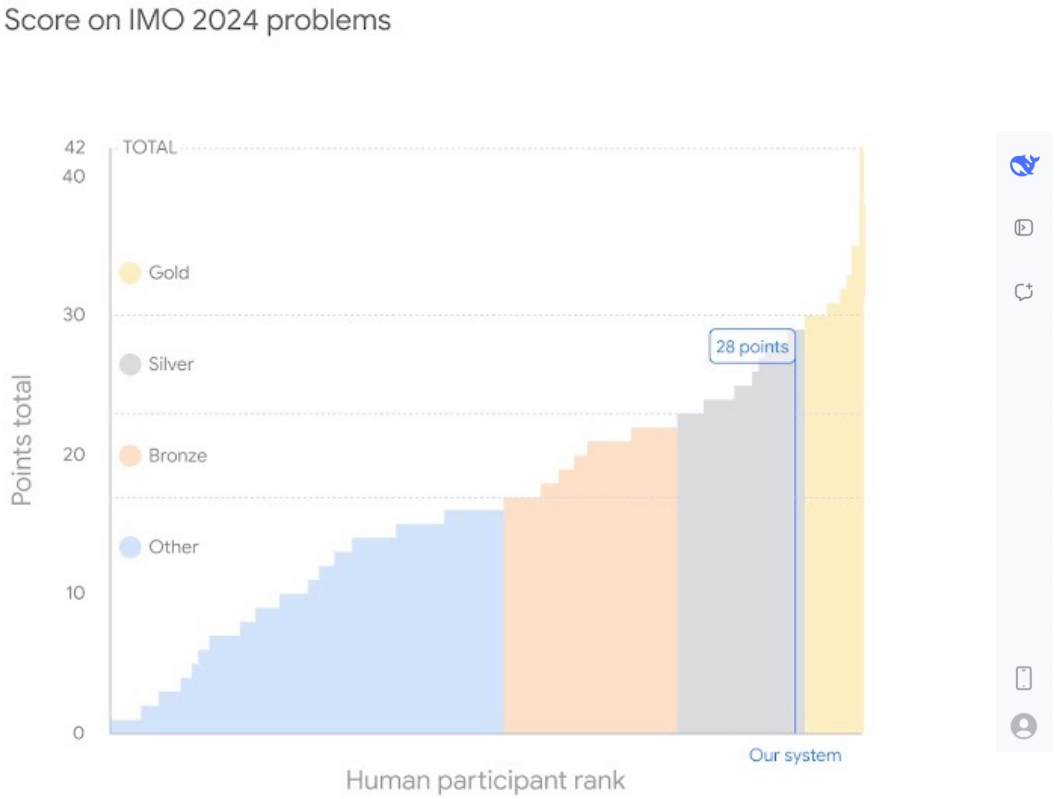
Language Models *RL*

- Use RL to improve performance of LLMs
- *LLM Alignment: to ensure its outputs are aligned with human intent, ethical principles, and task-specific requirements.*

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Success on Math

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



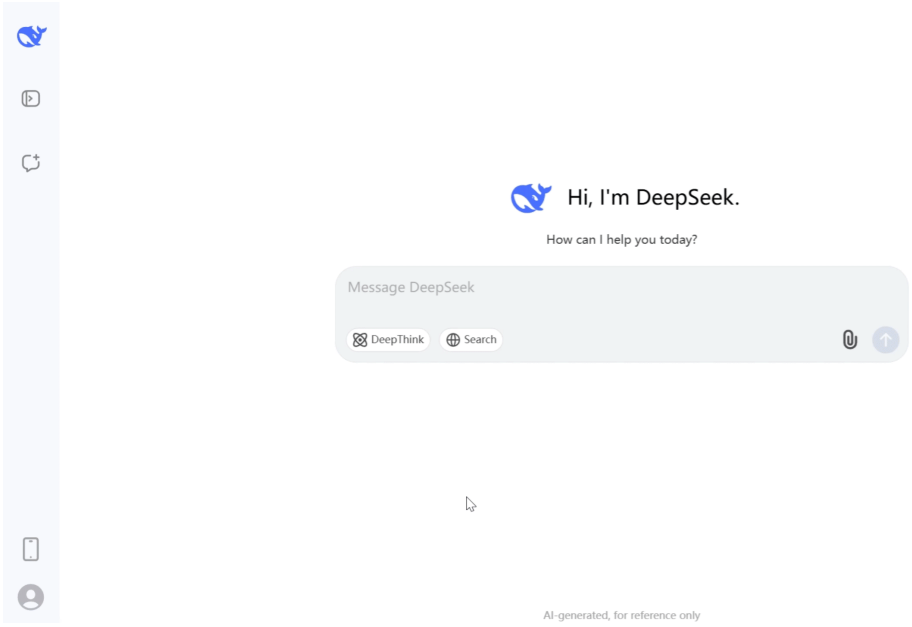


Image source: deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

Extending Such Success?

- Opportunities
  - Improvement of LLM abilities directly helps human
  - Inspirations from the RL world
- Challenges
  - Reward signals
  - Sufficient compute for simulation and optimization
  - Scalable & efficient algorithms: no silver bullet in RL
- Next Section:
  - RL, IRL, and RL in LLMs
  - LLM Alignment as Inverse RL --- RL w/ Data Driven Reward Function

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Extending Such Success?

- Opportunities
  - Improvement of LLM abilities directly helps human
  - Inspirations from the RL world
- Challenges
  - Reward signals
  - Sufficient compute for simulation and optimization
  - Scalable & efficient algorithms: no silver bullet in RL
- Next Section:
  - RL, IRL, and RL in LLMs
  - LLM Alignment as Inverse RL --- RL w/ Data Driven Reward Function

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Extending Such Success?

- Opportunities
  - Improvement of LLM abilities directly helps human
  - Inspirations from the RL world
- Challenges
  - Reward signals
  - Sufficient compute for simulation and optimization
  - Scalable & efficient algorithms: no silver bullet in RL
- Next Section:
  - RL, IRL, and RL in LLMs
  - LLM Alignment as Inverse RL --- RL w/ Data Driven Reward Function

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Extending Such Success?

- Opportunities
  - Improvement of LLM abilities directly helps human
  - Inspirations from the RL world
- Challenges
  - Reward signals
  - Sufficient compute for simulation and optimization
  - Scalable & efficient algorithms: no silver bullet in RL
- Next Section:
  - RL, IRL, and RL in LLMs
  - LLM Alignment as Inverse RL --- RL w/ Data Driven Reward Function

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

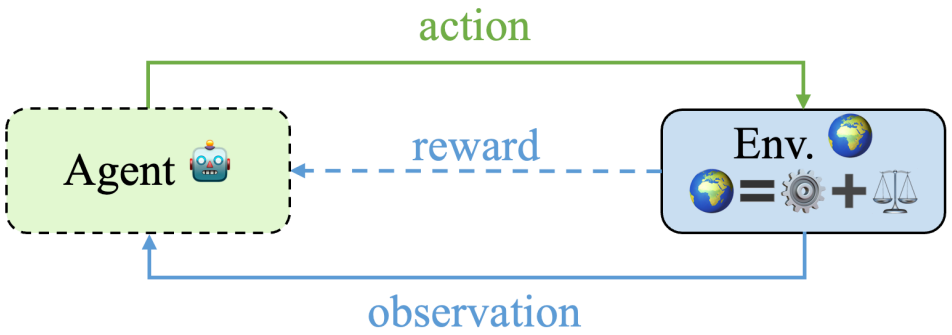
Part 1:

Reinforcement Learning: Forward and Inverse

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL: Learning through Interactions

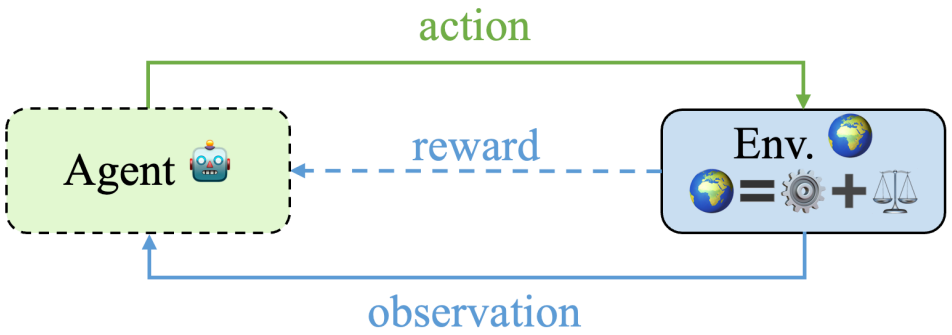
- Learning a policy to maximize the long-term return by interacting with the environment.



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

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

RL: Learning through Interactions

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

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

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

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

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

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

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*

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***

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL: Learning through Interactions

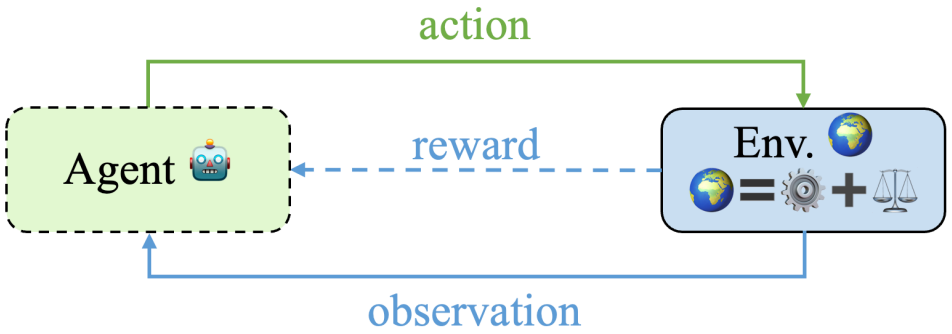
- *Some* RL algorithms can be better than others in *some* tasks
- There is ***no silver bullet in RL***
- e.g.,

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?

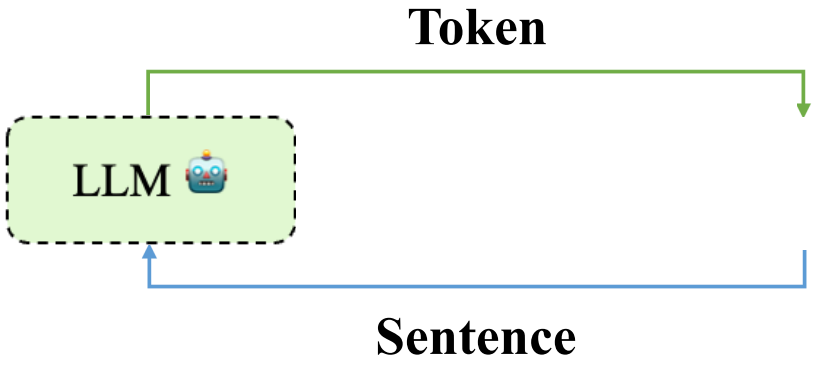
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Generation as an MDP?

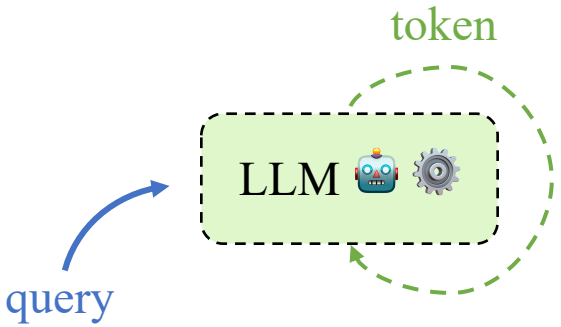
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- S: Current sentence
- A: Tokens (or their combinations)



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Generation as an MDP?

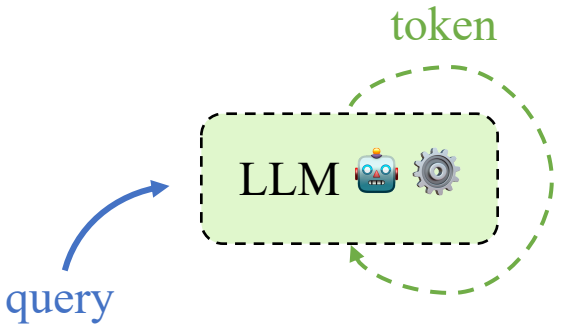
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- S: Current sentence
- A: Tokens (or their combinations)
- P: Concatenation of tokens
- ρ_0: Prompt/Query distribution



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

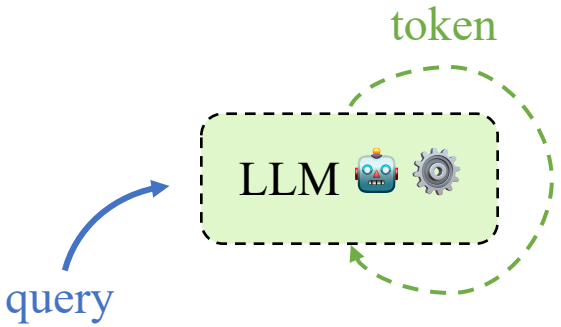
LLM Generation as an MDP\R

- 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

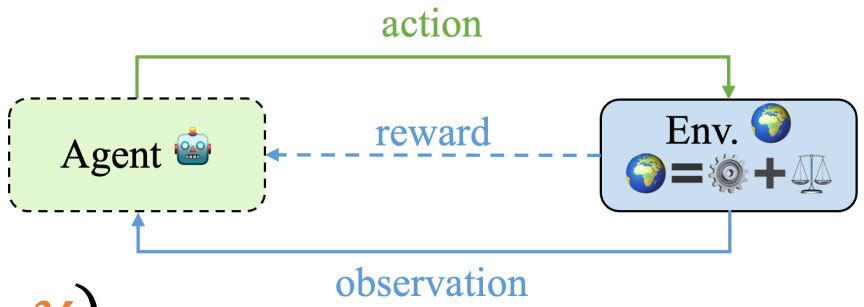
- Markov Decision Process: $\mathcal{M} = (S, A, P, \rho_0, R, \gamma)$
- S: Current sentence
- A: Tokens (or their combinations)
- P: Concatenation of tokens
- ρ_0: Prompt/Query distribution
- ? R: *(Data-Driven)*
- γ: ≤ 1 (e.g., =1: correct is enough / <1: correct in short answer)



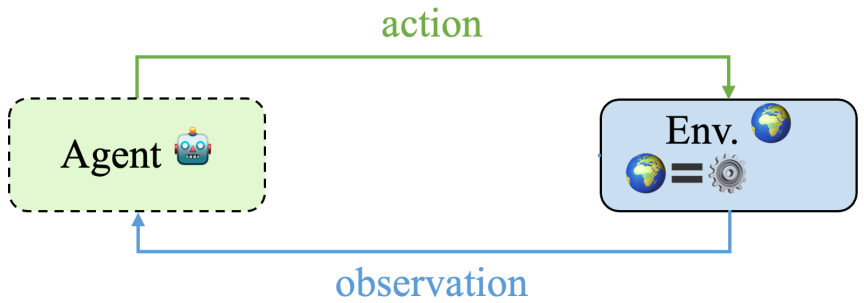
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

What is MDP\R

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



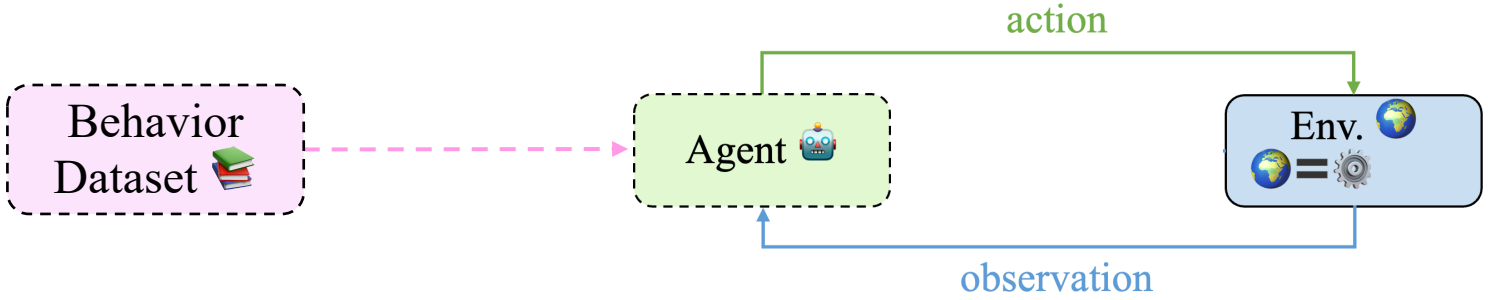
- In MDP\Rs (S, A, P, ρ_0, γ), how (what) to learn?



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]

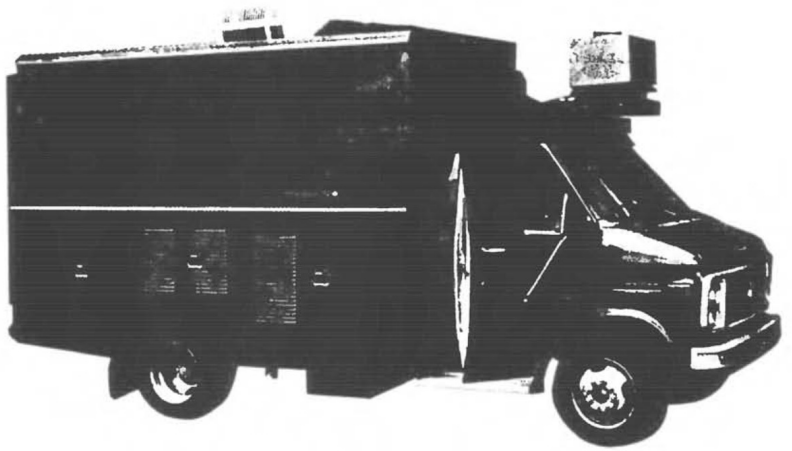


Figure 3: NAVLAB, the CMU autonomous navigation test vehicle.

Examples

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

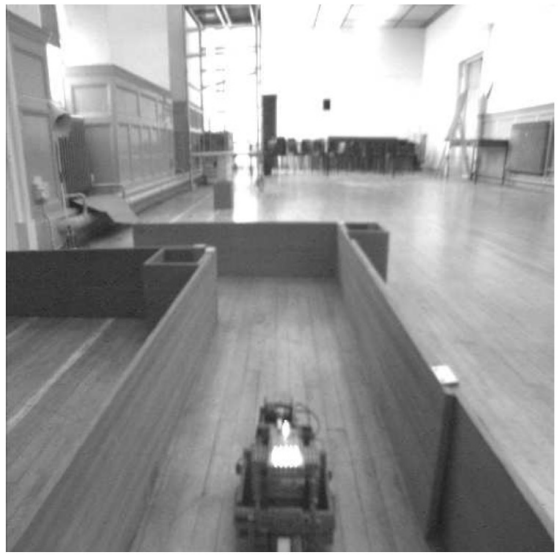


Fig. 3.: View of the teacher and part of the maze as seen by Ben Hope.

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]

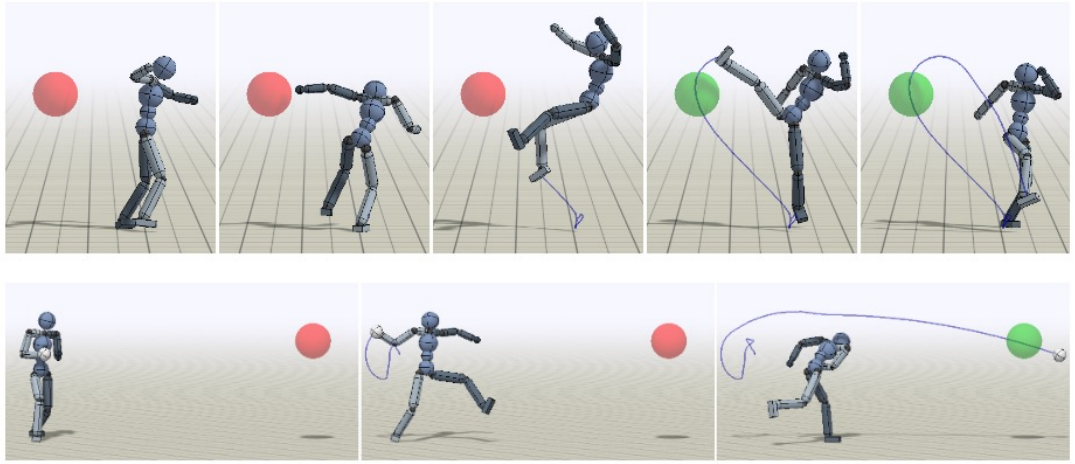


Fig. 7. **Top:** Spinkick policy trained to strike a target with the character's right foot. **Bottom:** Baseball pitch policy trained to throw a ball to a target.

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]

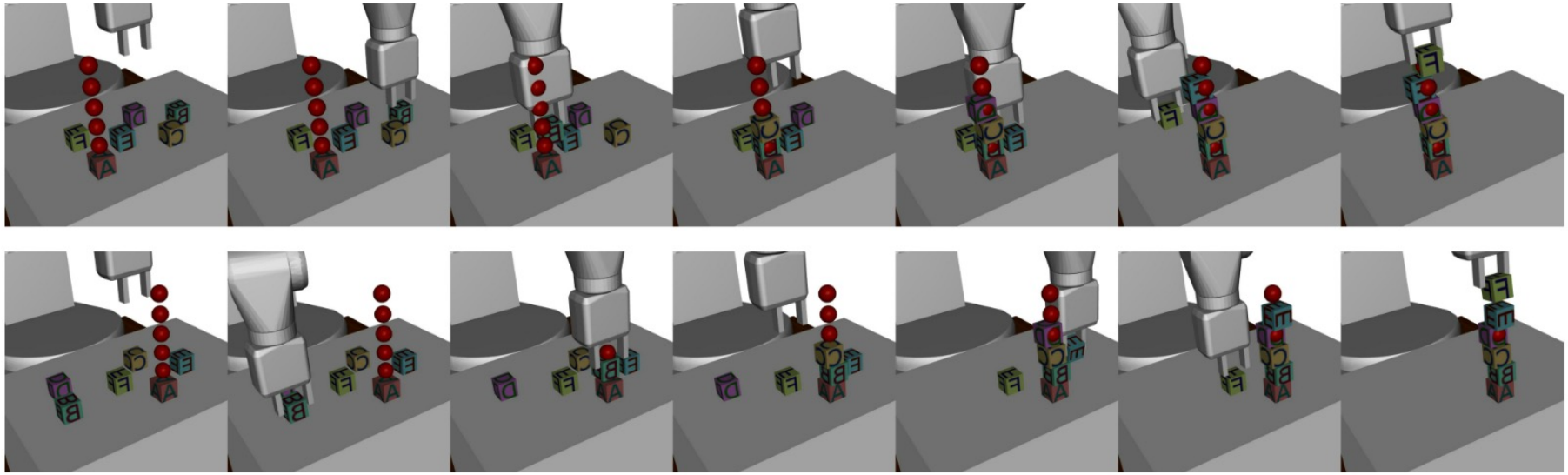
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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

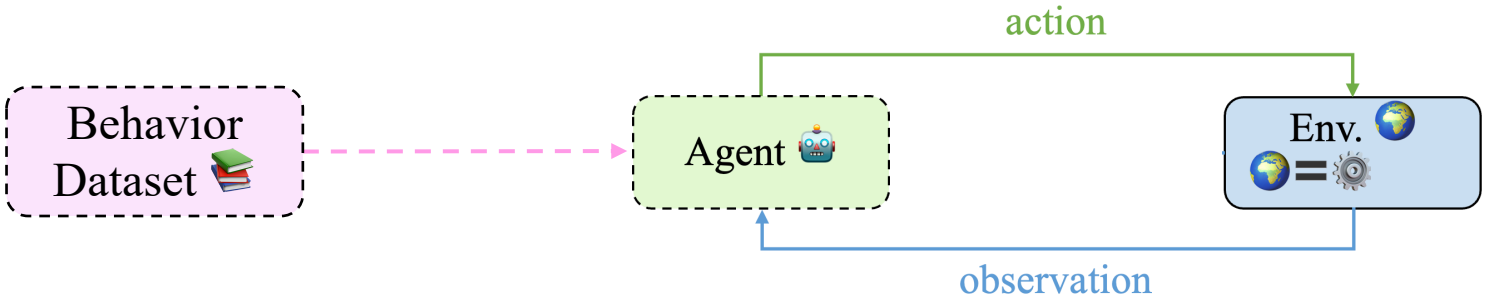


- 2. Reward signal is too sparse (e.g., win a game of Go/StarCraft/Dota2)
  - AlphaGo/AlphaStar/OpenAI Five
  - Robotics Control [Nair et al., 2017]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Methods for Learning from Behavior

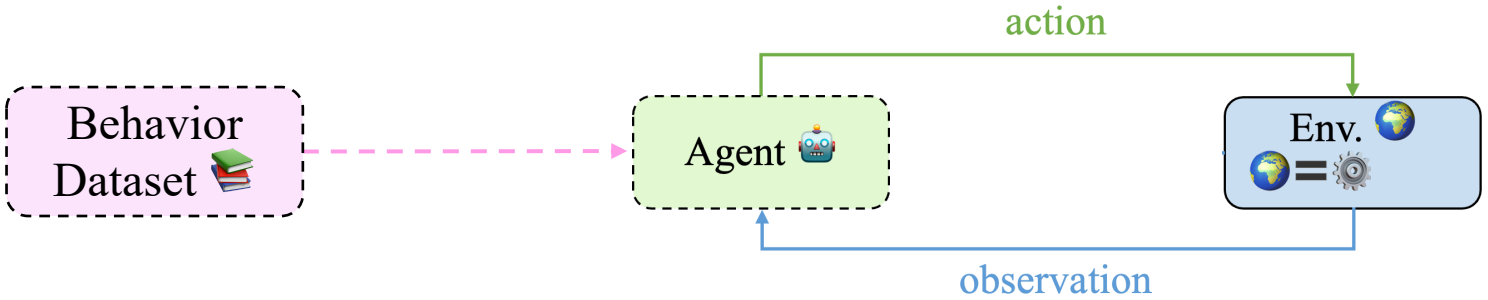
- Learning from a *behavior dataset*
  - Imitation Learning: recover π^* given behavior generated by π^*



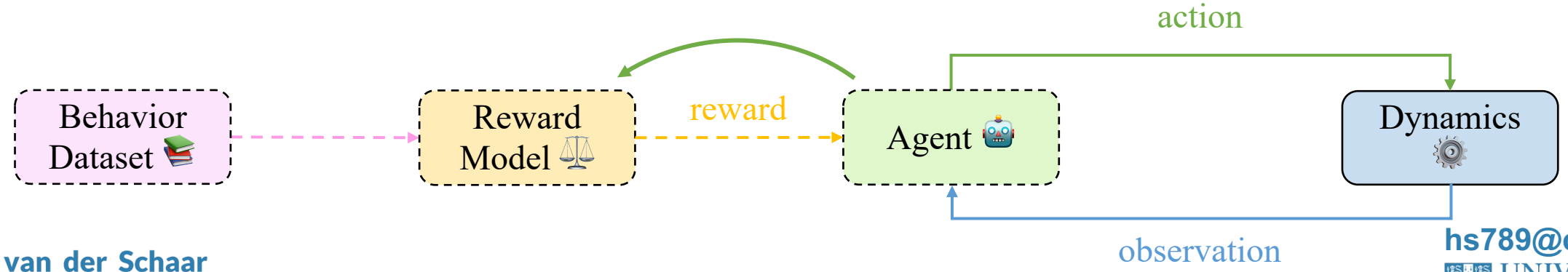
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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^*

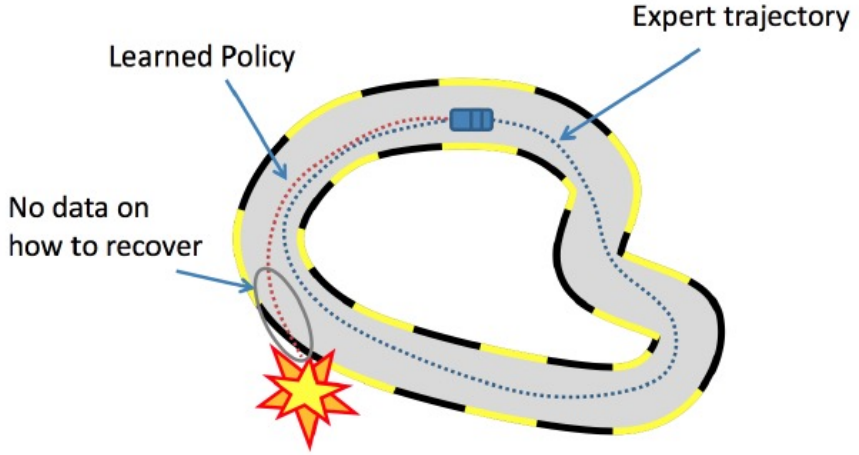
  - IL 1. Behavior Clone
    [Hayes & Demiris, 1994] [Pomerleau, 1988]

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]

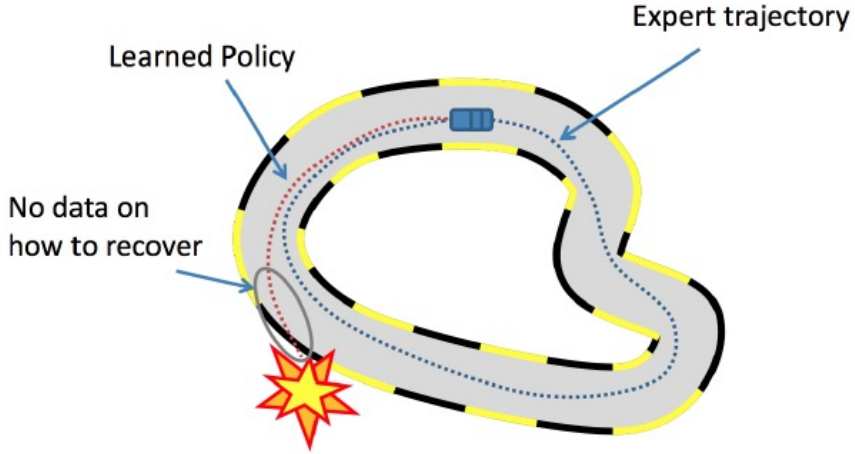


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Imitation Learning

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

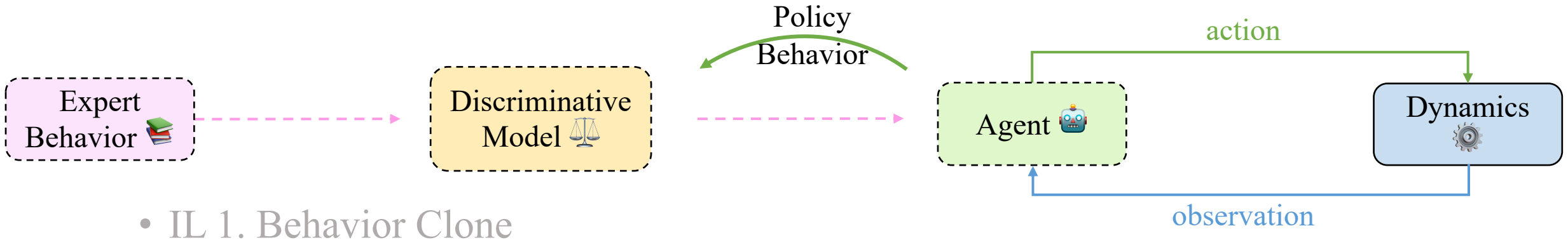
- IL 1. Behavior Clone
  [Hayes & Demiris, 1994; Pomerleau, 1988]
- IL 2. Expert Engagement
  [Ross et al. 2011; Li et al., 2022; Peng et al. 2023]



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Imitation Learning

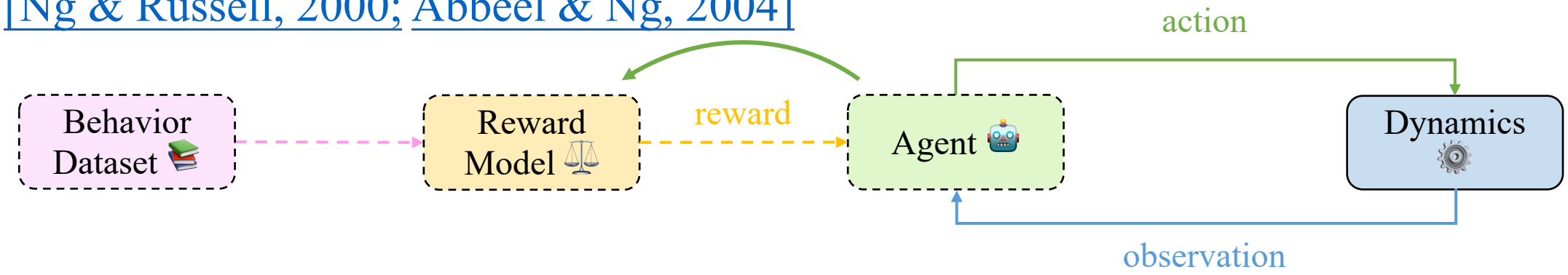
- Imitation Learning: recover π^* given behavior generated by π^*



  - 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]
  - IL 3. Adversarial Imitation
    [Ho & Ermon, 2016]

Inverse Reinforcement Learning

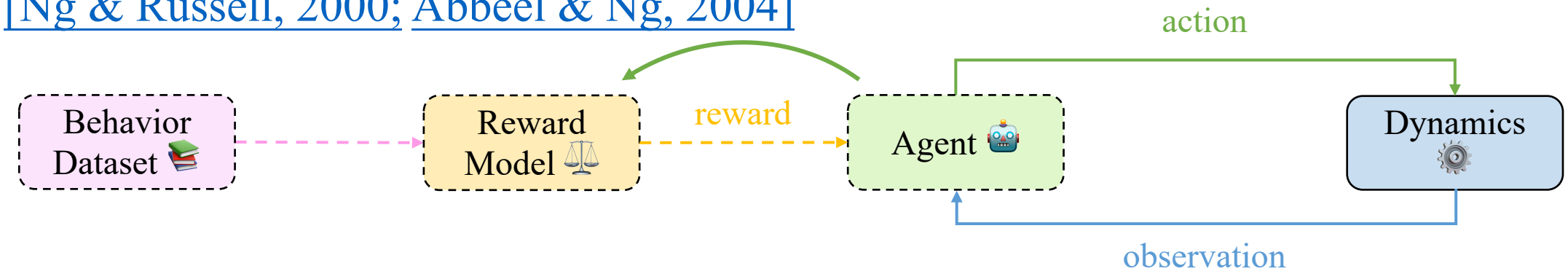
- Inverse RL: recover R that induces π^* given behavior generated by π^*
  [Ng & Russell, 2000; Abbeel & Ng, 2004]



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]

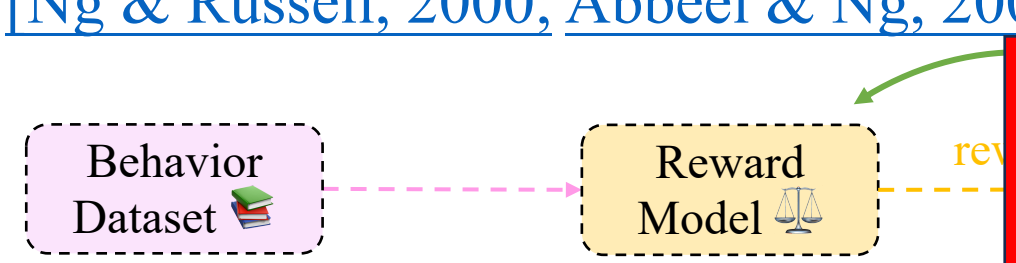


  - Max-Ent IRL [Ziebart et al., 2008]
  - Adversarial IRL [Fu et al., 2017]
  - T-REX [Brown et al., 2019]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

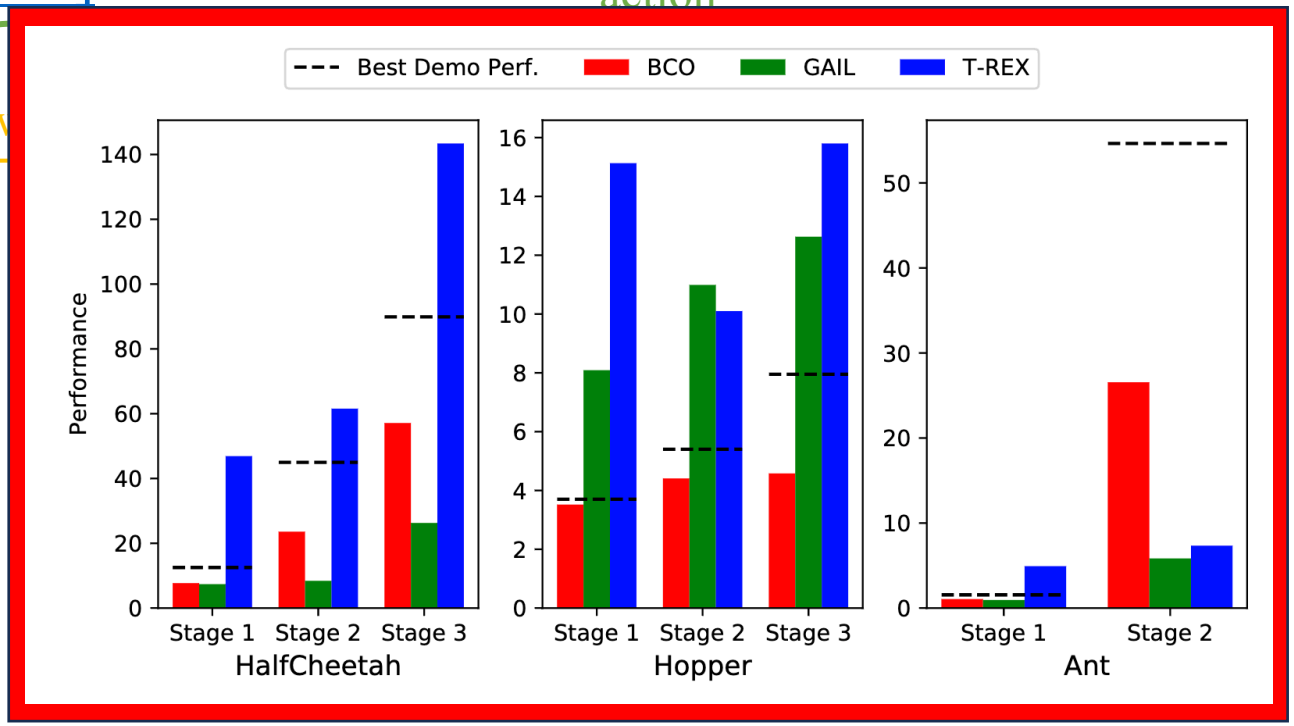
Inverse Reinforcement Learning

- 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]

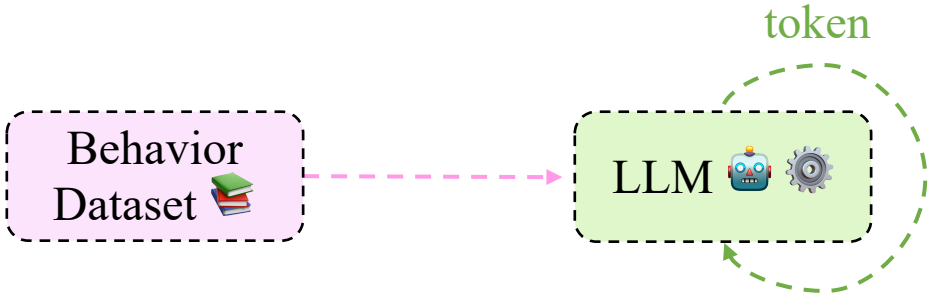
Can outperform demonstration



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Optimization via Imitation

- LLMs as language imitators

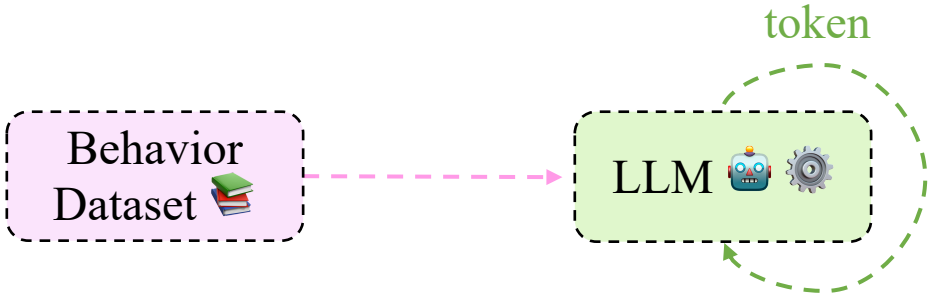


- Pre-train: large scale behavior clone
  [Obtain (strong) ability of understanding]

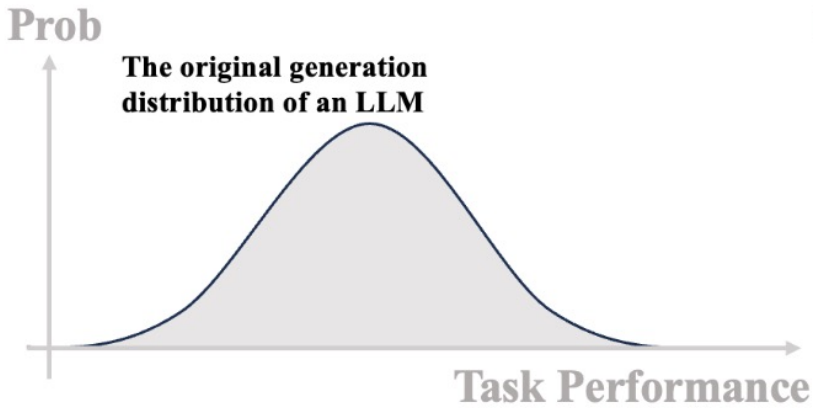
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Optimization via Imitation

- LLMs as language imitators



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

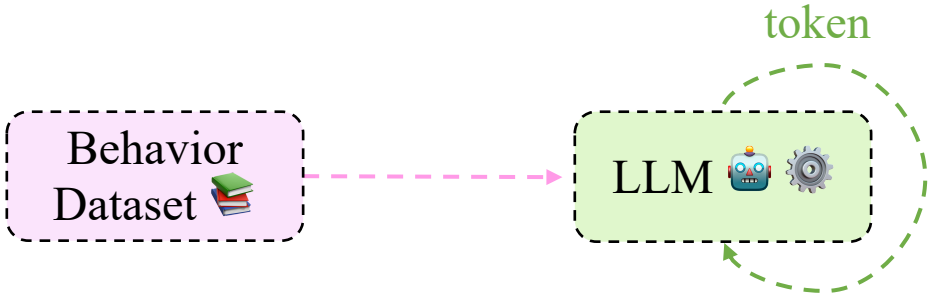


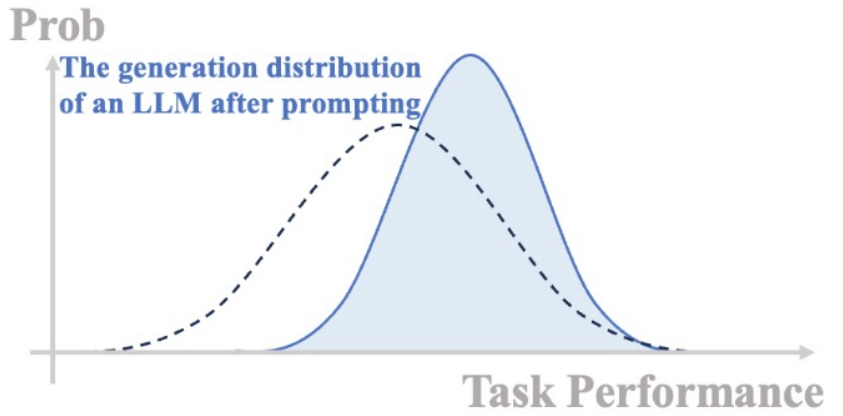
(1) LLM *Can* do Any Task as a *Universal Sampler*

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLM Optimization via Imitation

- LLMs as language imitators





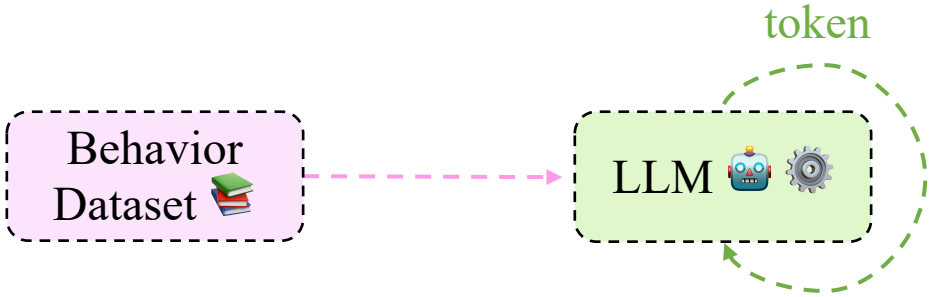
(2) Prompting *Can* Improve Performance by shifting the generation

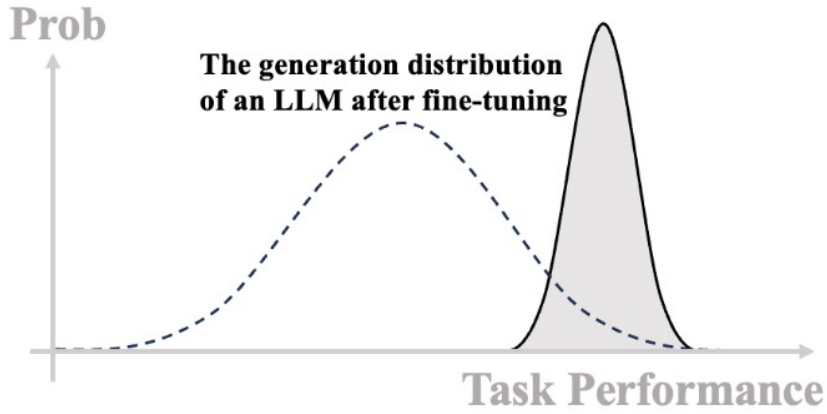
- Pre-train: large scale behavior clone
  [Obtain (strong) ability of understanding]
- Post-train/alignment: optimization on a specific task
  - Smart prompting strategy [Kojima et al., 2022]

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?

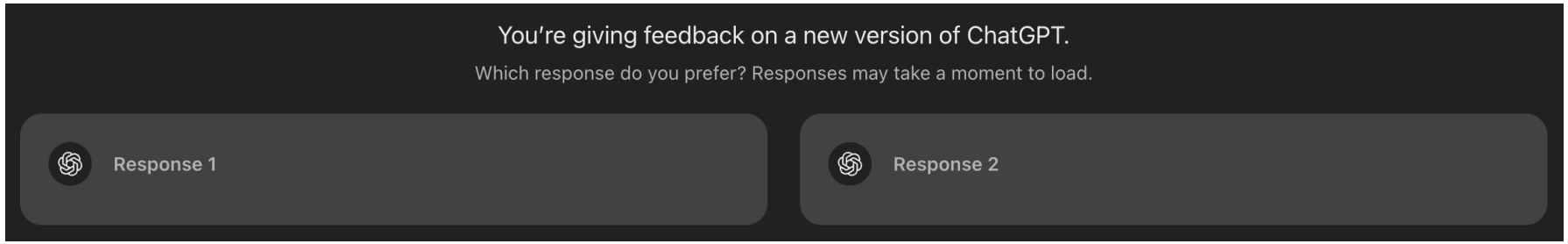
- Preference feedback can be more scalable & practical

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Why Do We Need Inverse RL?

- Preference feedback can be more scalable & practical

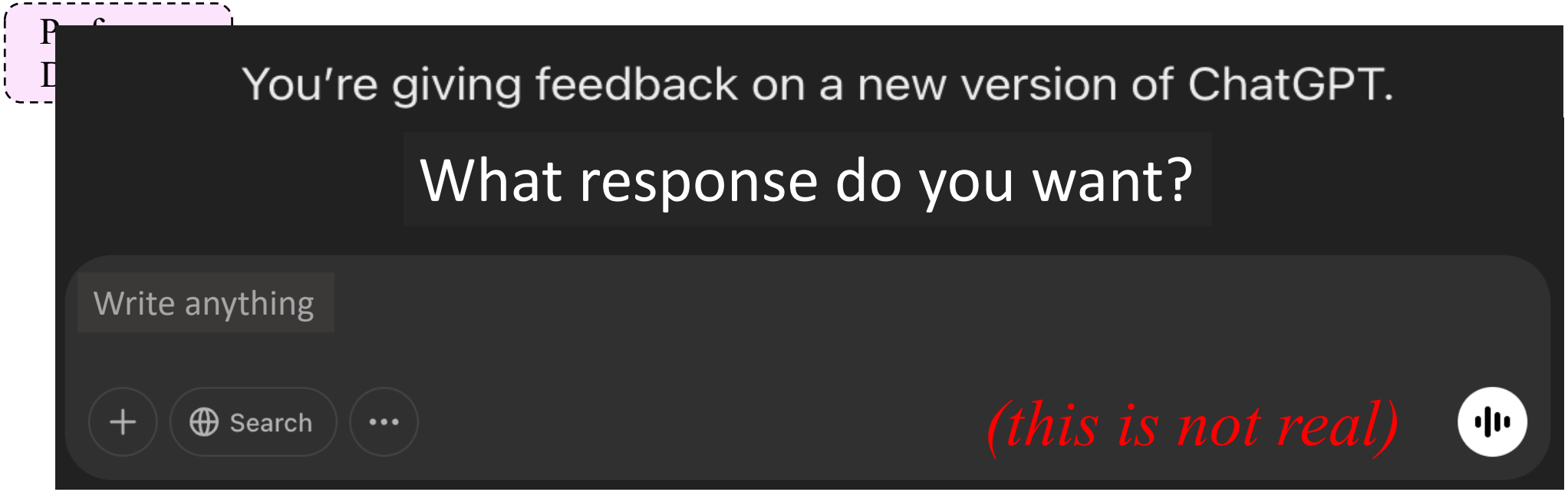
Preference Dataset 📚



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

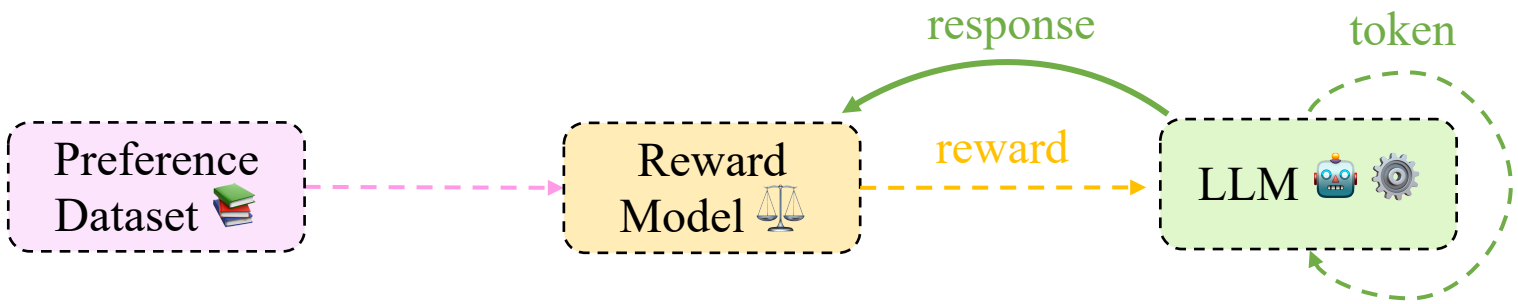
Why Do We Need Inverse RL?

- Preference feedback can be more scalable & practical



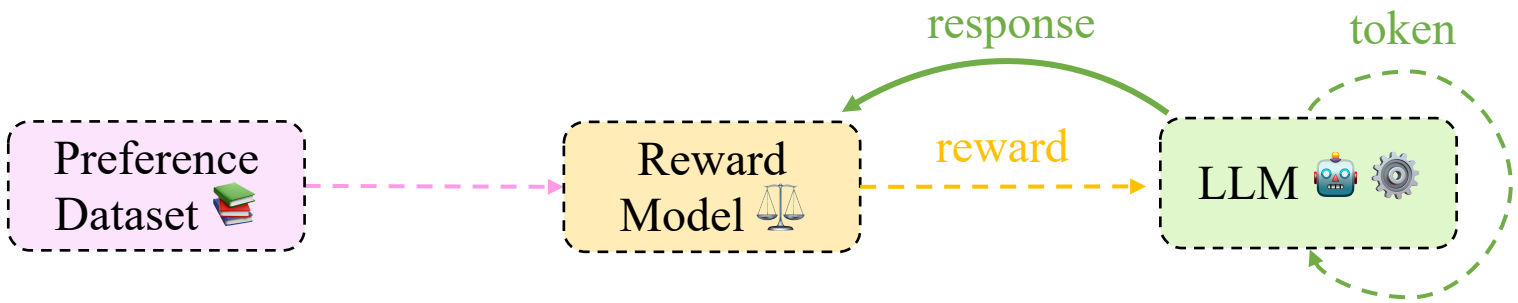
Why Do We Need Inverse RL?

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

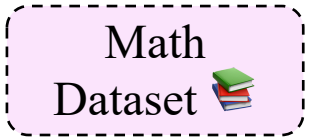


Why Do We Need Inverse RL?

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



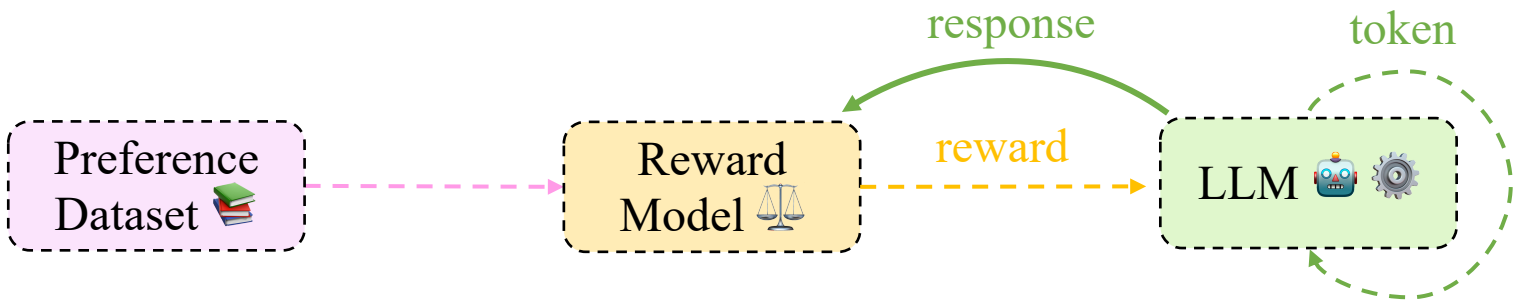
- Math: find a more generalizable reasoning path toward correct answers



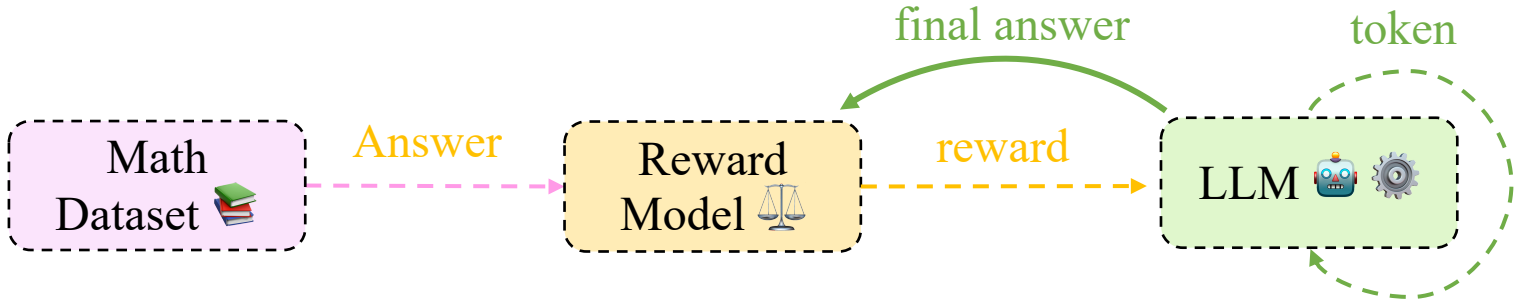
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Why Do We Need Inverse RL?

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

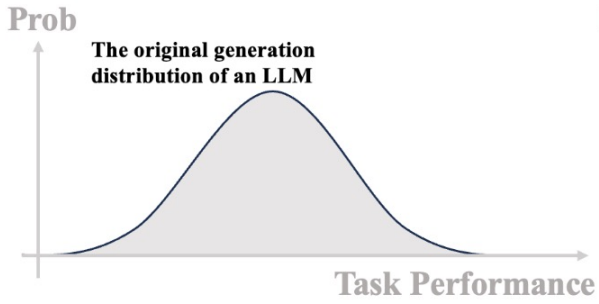


- Math: find a more generalizable reasoning path toward correct answers

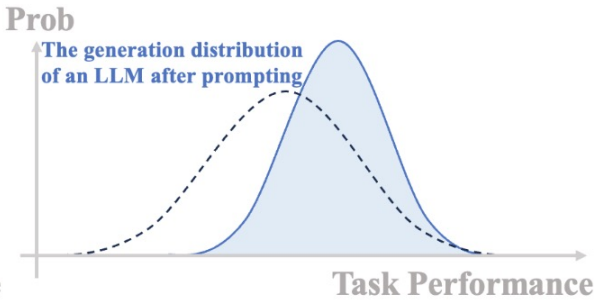


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

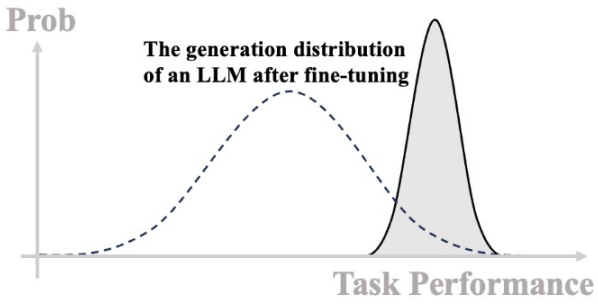
RMs Enable Test-Time Optimization [Sun et al., 2024]



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



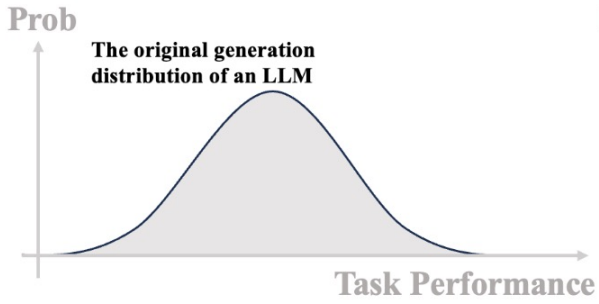
(2) Prompting *Can* Improve Performance by shifting the generation



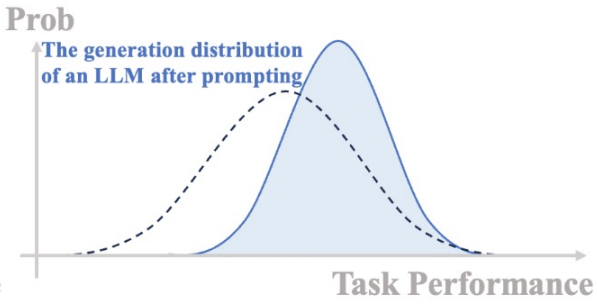
(3) Fine-Tuning *Can* Improve Performance by shifting the generation

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

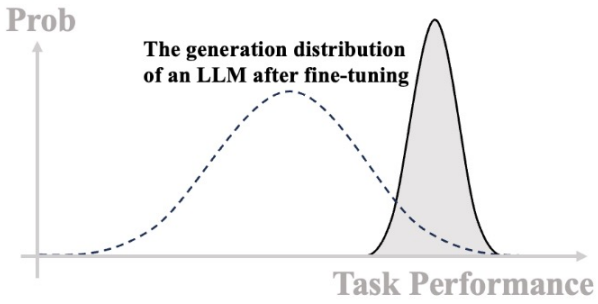
RMs Enable Test-Time Optimization [Sun et al., 2024]



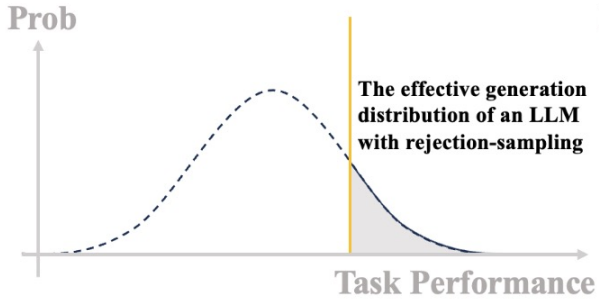
(1) LLM *Can* do Any Task as a *Universal Sampler*



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



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



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

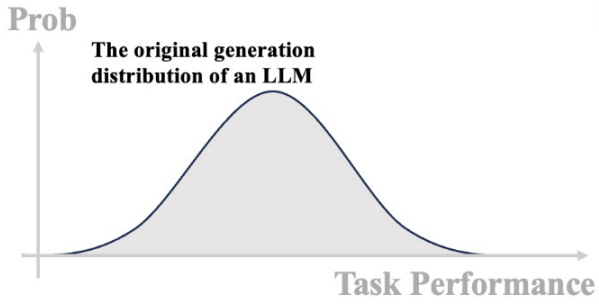
Reward Models
Enable Test-time Optimization

van_der_Schaar \ LAB

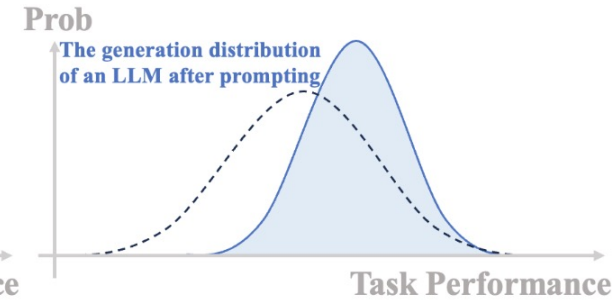
sites.google.com/view/irl-llm

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

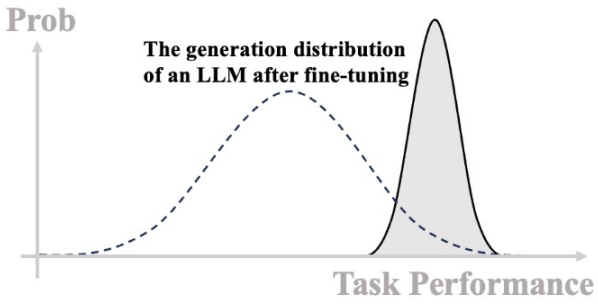
RMs Enable Test-Time Optimization [Sun et al., 2024]



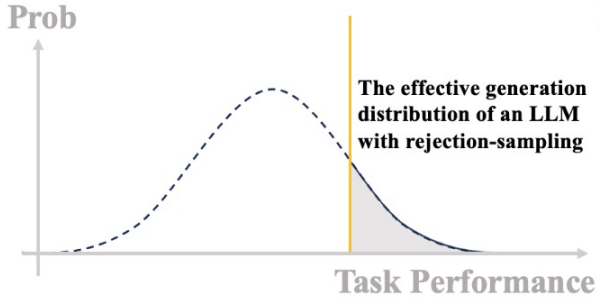
(1) LLM *Can* do Any Task as a *Universal Sampler*



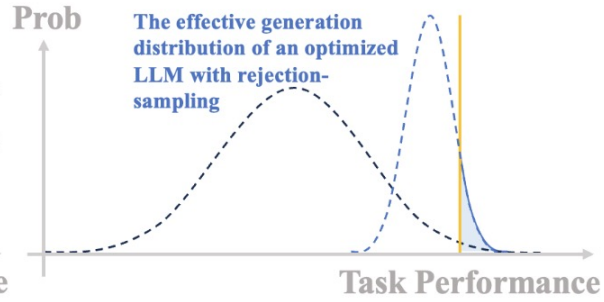
(2) Prompting *Can* Improve Performance by shifting the generation



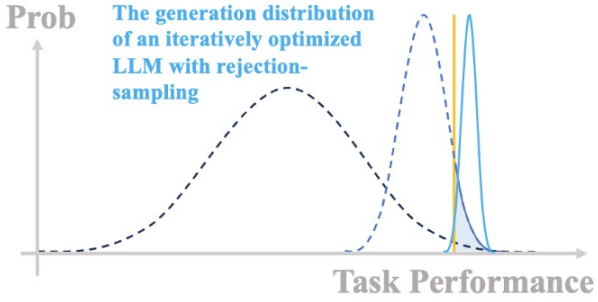
(3) Fine-Tuning *Can* Improve Performance by shifting the generation



(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 2:

Learning Reward Models from Data

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Learning Reward Models from Data

RMs from different data types

RM from Binary (e.g., Math Reasoning)

RM from Preference (e.g., Classical RLHF)

RM from Demonstration (e.g., Expert Data)

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…

(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:

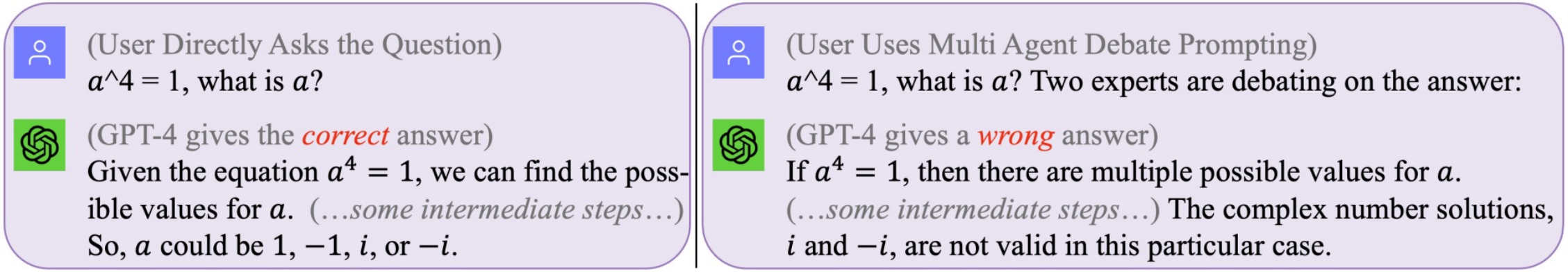
(GPT-4 gives a *wrong* answer)
If $a^4 = 1$, then there are multiple possible values for a. (…*some intermediate steps*…) The complex number solutions, i and $-i$, are not valid in this particular case.

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…



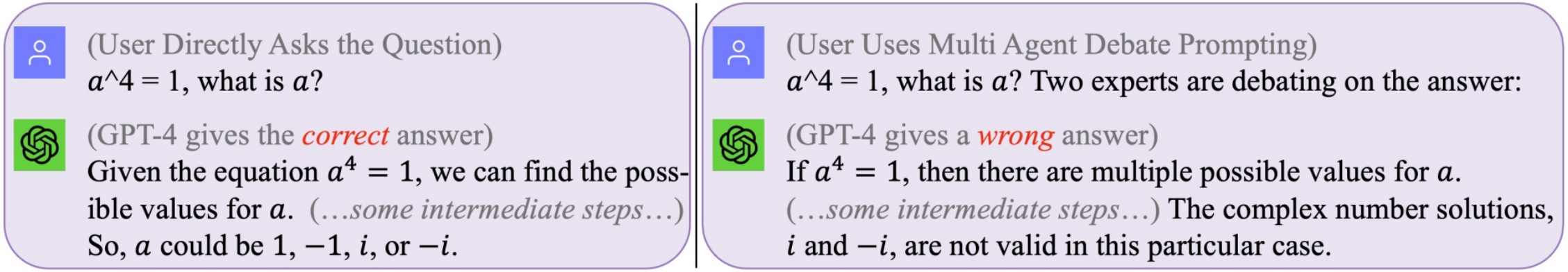
- Automatic prompt engineering using RL?

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

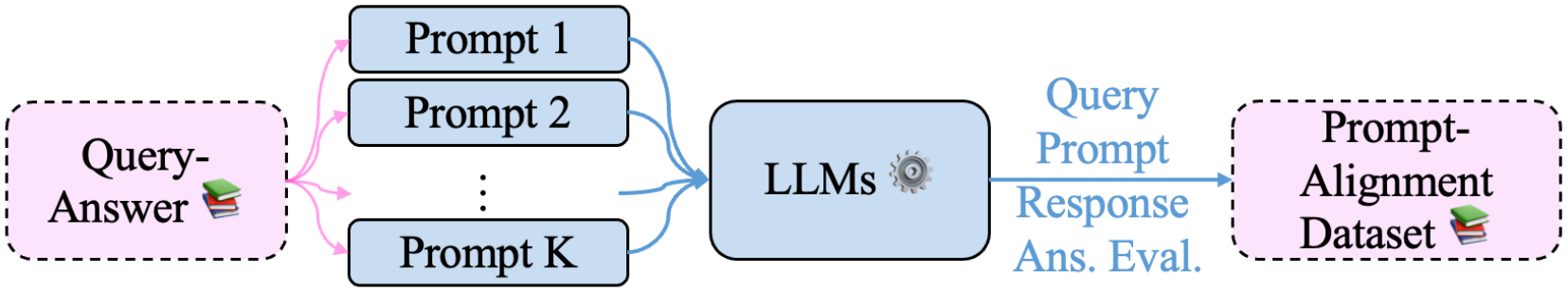
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

Prompt-OIRL [Sun et al., 2024]

- Learning from demonstrative behaviors can be more efficient!

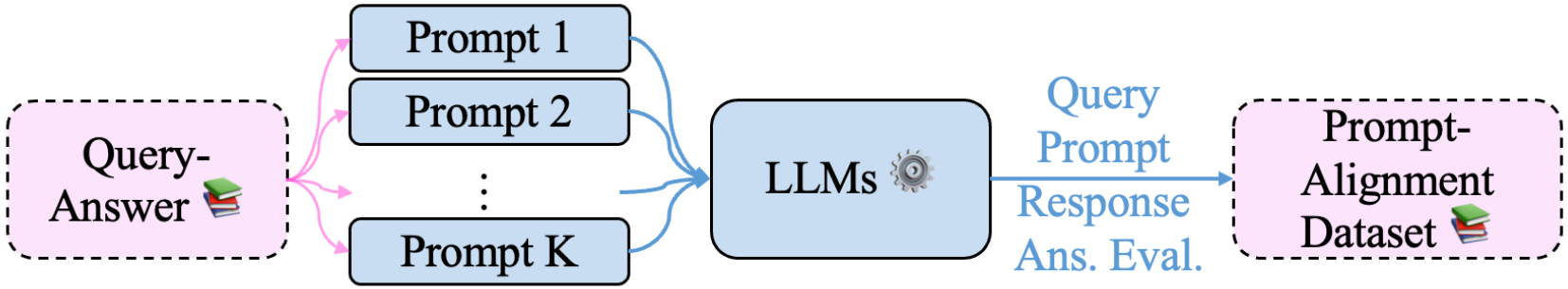
- Prompt Optimization as Inverse RL: learning from expert prompts



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

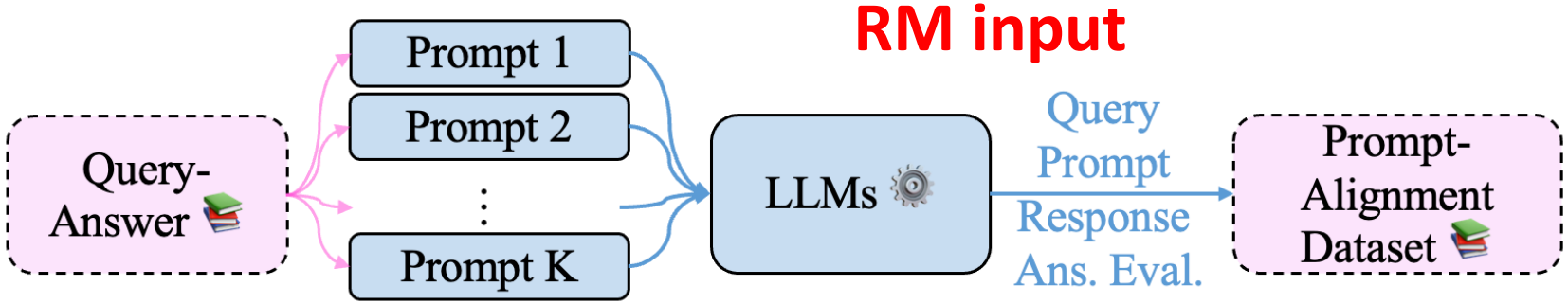
- Prompt-alignment dataset: [Query, Prompt, Response, Correctness]



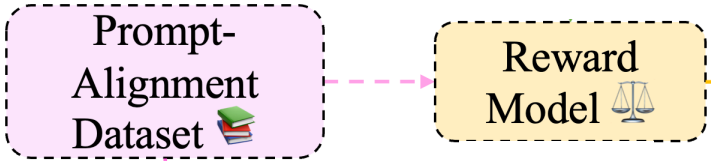
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

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



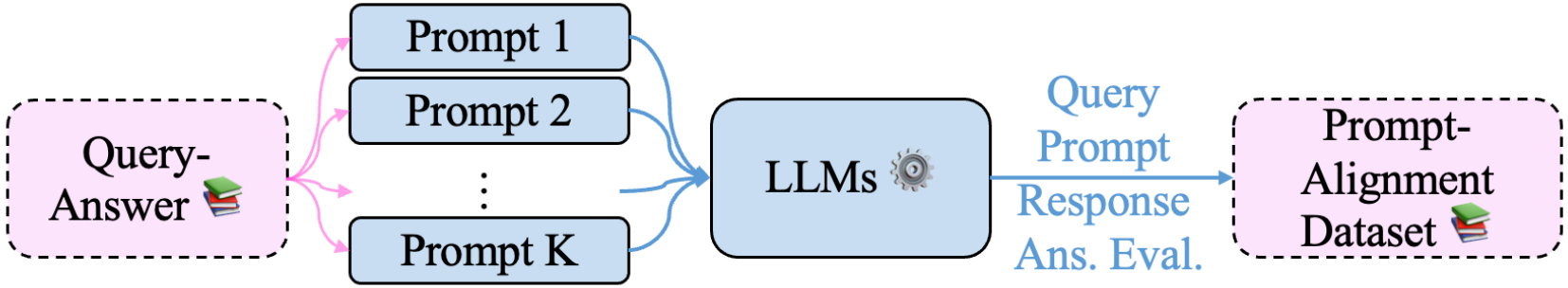
- Inverse RL:
  - (training) Reward Model



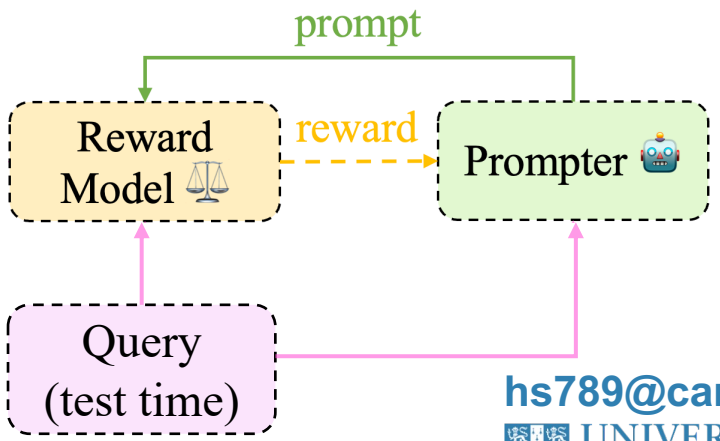
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

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

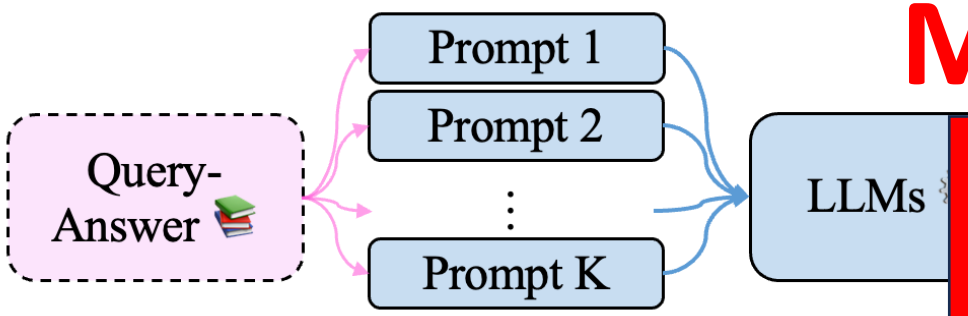


- Inverse RL:
  - (training) Reward Model
  - (test-time) Prompt Optimization
    select the *best* prompt for each query using RM

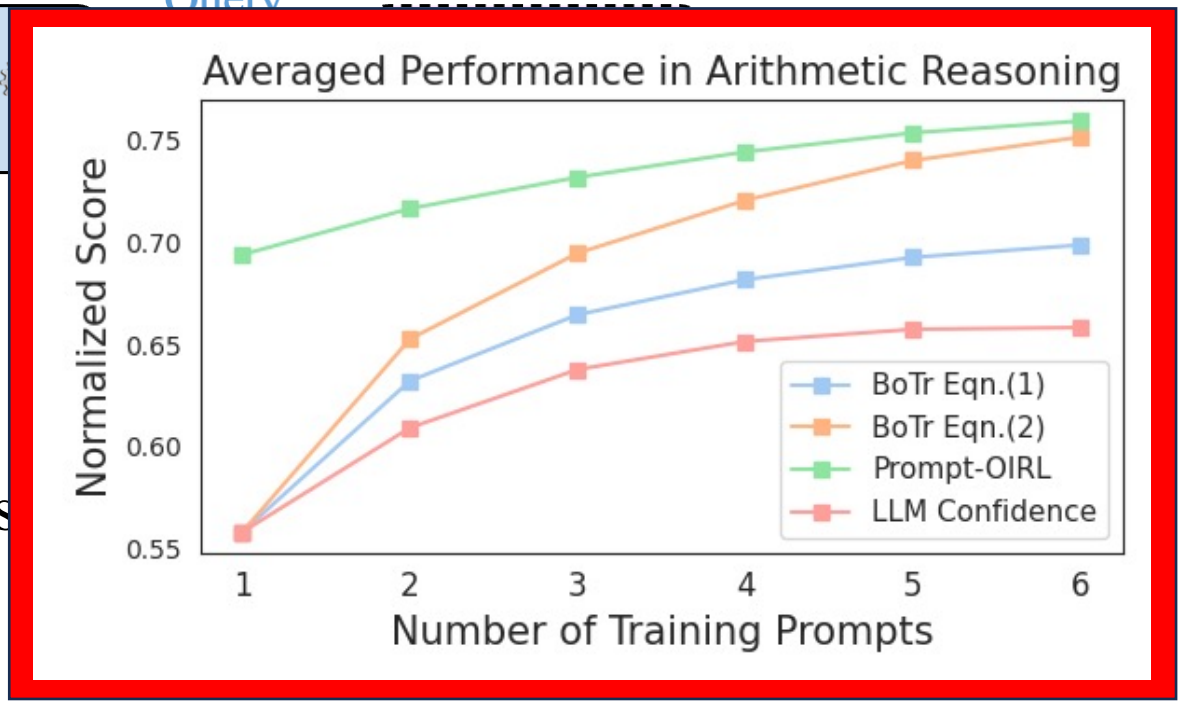


RM from Binary Data: IRL for *Prompt Optimization*

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



Math Reasoning Ability



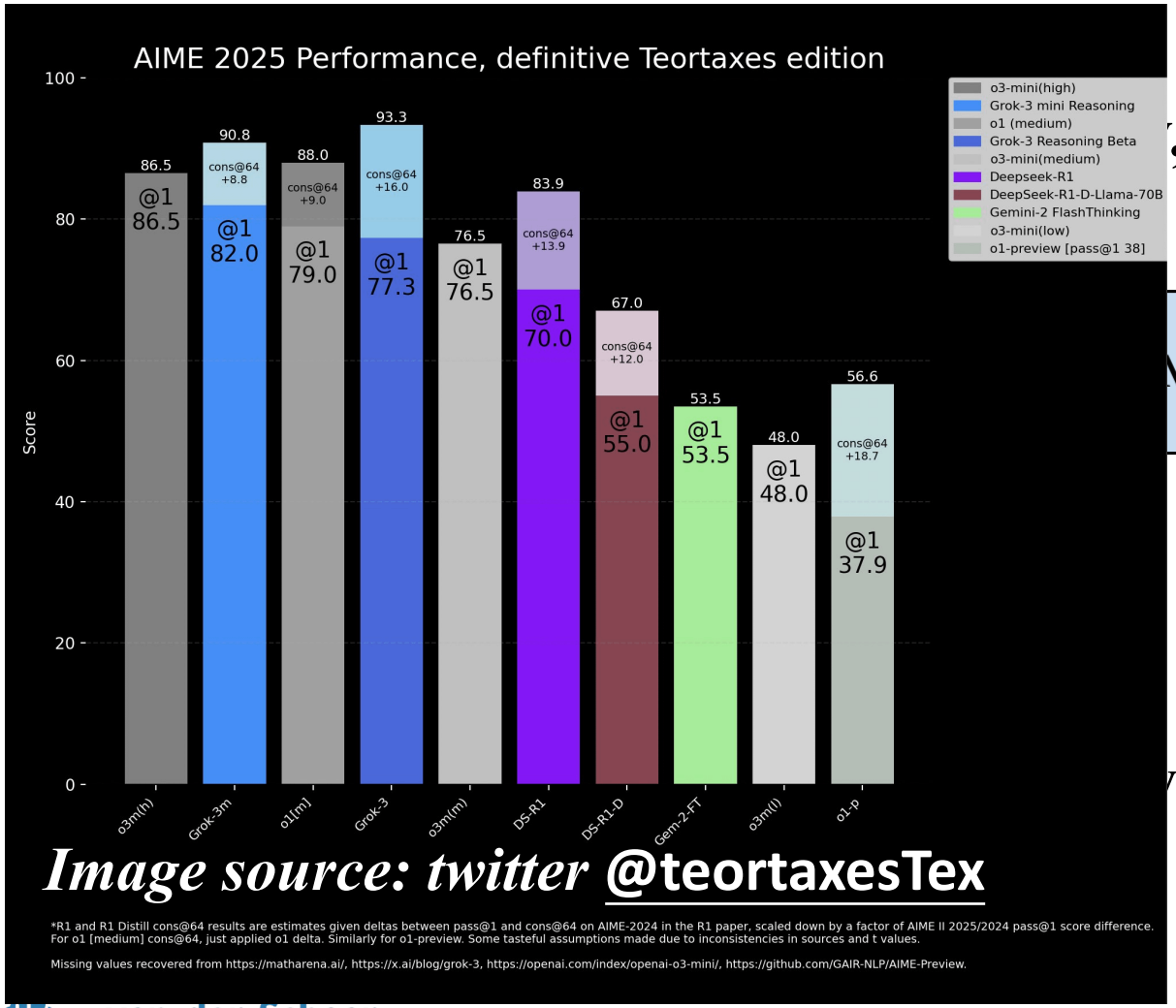
- Inverse RL:
  - (training) Reward Model
  - (test-time) Prompt Optimization
    select the *best* prompt for each query us

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

No need for 64 forward passes!

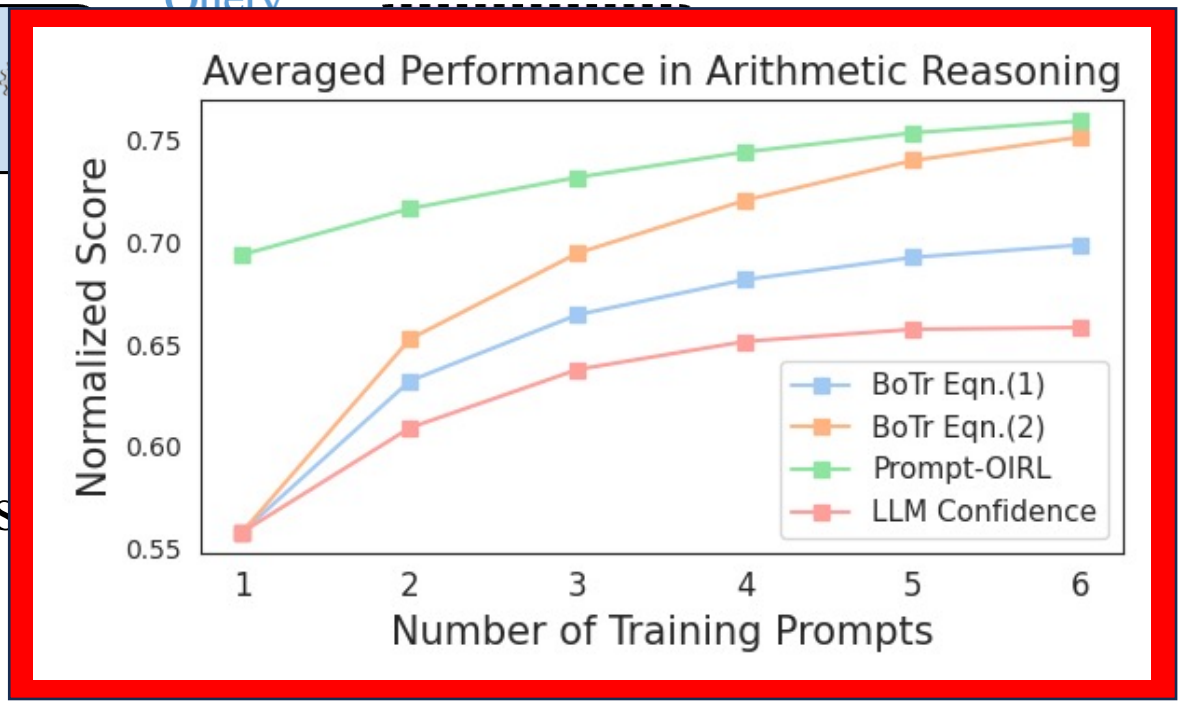
Pick the best using RM and do one single pass ☺

RM from Binary Data: IRL for *Prompt Optimization*



, Prompt, Response, Correctness]

Math Reasoning Ability



van_der_Schaar \ LAB

hs789@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

sites.google.com/view/irl-llm

Learning Reward Models from Data

RMs from different data types

RM from Binary (e.g., Math Reasoning)

RM from Preference (e.g., Classical RLHF)

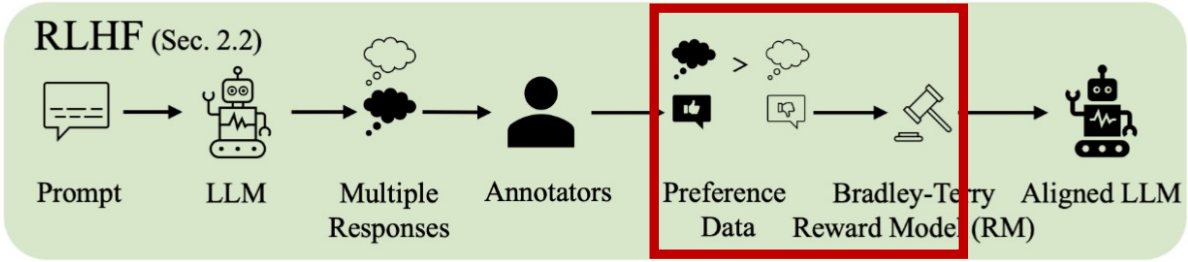
RM from Demonstration (e.g., Expert Data)

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Preference-Reward Model Foundation Paper [Sun et al., 2025]

RM from Preference: Back to Ranking Theory

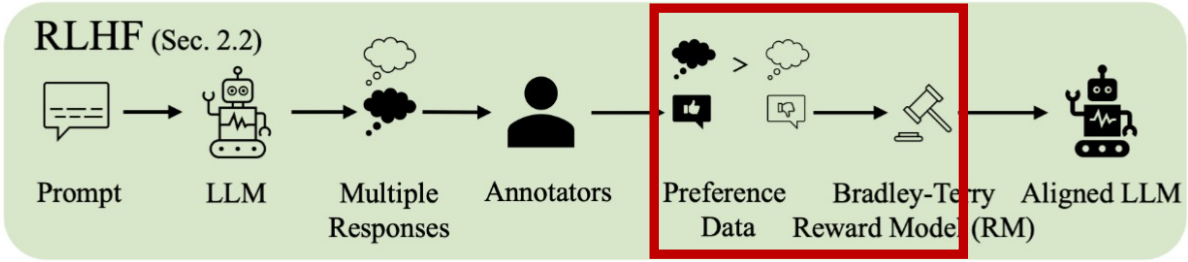
- How to build reward model?
  - RLHF: “use the Bradley-Terry Model”
  - But why?



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- How to build reward model?
  - RLHF: "use the Bradley-Terry Model"
  - But why?



- 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,

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

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Preference-Reward Model Foundation Paper [Sun et al., 2025]

RM from Preference: Back to Ranking Theory

- 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 150 models, 2M competitions, each LLM plays ~10,000 games

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Preference-Reward Model Foundation Paper [Sun et al., 2025]

RM from Preference: Back to Ranking Theory

- 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 150 models, 2M competitions, each LLM plays ~10,000 games

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes
1	1	Gemini-2.0-Flash-Thinking-Exp-01-21	1384	+6/-9	10022
1	1	Gemini-2.0-Pro-Exp-02-05	1379	+7/-6	7853
3	1	ChatGPT-4o-latest (2024-11-20)	1365	+4/-3	38148
3	1	DeepSeek-R1	1361	+8/-7	4193
3	7	Gemini-2.0-Flash-001	1357	+8/-7	5576
4	1	o1-2024-12-17	1352	+6/-6	12021
7	5	o1-preview	1335	+4/-3	33174
7	7	Qwen2.5-Max	1332	+12/-11	3394
8	9	DeepSeek-V3	1316	+5/-5	16583
9	9	Gemini-2.0-Flash-Lite-Preview	1306	+9/-8	5083
9	12	GLM-4-Plus-0111	1305	+11/-10	2791
9	13	Step-2-16K-Exp	1304	+8/-9	5128
10	13	o1-mini	1305	+3/-4	52726
10	9	Gemini-1.5-Pro-002	1302	+3/-4	49653

Image source: lmsys-llm-arena, screenshot date: Feb-5-2025

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Preference-Reward Model Foundation Paper [Sun et al., 2025]

RM from Preference: Back to Ranking Theory

- 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 150 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 $\boldsymbol{N \log N}$

Preference-Reward Model Foundation Paper [Sun et al., 2025]

RM from Preference: Back to Ranking Theory

- 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 150 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 $\boldsymbol{N\log N}$

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

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- Why does BT model work?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- Why does BT model work?
  - An analogy:

    knowing the results of 5 groups of 1 on 1 running matches between 10 people.

    predict who is the best runner in **another** 100 people.

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- Why does BT model work?
  - An analogy:

    knowing the results of 5 groups of 1 on 1 running matches between 10 people.

    predict who is the best runner in **another** 100 people.

  - It's impossible, unless…

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- Why does BT model work?
  - An analogy:

    knowing the results of 5 groups of 1 on 1 running matches between 10 people.

    predict who is the best runner in **another** 100 people.

  - It's impossible, unless we have a stopwatch!

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- 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

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- 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?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- 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
  - Is it necessary in RLHF?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- 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
  - Is it necessary in RLHF?

    NO, we only need *order consistency*

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Order Consistency

- In test-time optimization, we use RM to pick the best of possible choices.
- e.g.,

Q-A Pairs	Q + Answer 1	Q + Answer 2	Q + Answer 3	Q + Answer 4
True Reward	9.8	9.0	7.0	6.0
Reward Model 1	1.0	0.9	0.8	0.7
Reward Model 2	100	99	80	50
Reward Model 3	9.0	9.8	7.0	6.0

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Order Consistency

- In test-time optimization, we use RM to pick the best of possible choices.
- e.g.,

Q-A Pairs	Q + Answer 1	Q + Answer 2	Q + Answer 3	Q + Answer 4
True Reward	9.8	9.0	7.0	6.0
Reward Model 1	1.0	0.9	0.8	0.7
Reward Model 2	100	99	80	50
Reward Model 3	9.0	9.8	7.0	6.0

- RM 1 = RM 2 > RM 3

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Order Consistency

- In test-time optimization, we use RM to pick the best of possible choices.
- e.g.,

Q-A Pairs	Q + Answer 1	Q + Answer 2	Q + Answer 3	Q + Answer 4
True Reward	9.8	9.0	7.0	6.0
Reward Model 1	1.0	0.9	0.8	0.7
Reward Model 2	100	99	80	50
Reward Model 3	9.0	9.8	7.0	6.0

- Spearman Ranking Correlation

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Order Consistency

- In test-time optimization, we use RM to pick the best of possible choices.
- e.g.,

Q-A Pairs	Q + Answer 1	Q + Answer 2	Q + Answer 3	Q + Answer 4
True Reward	9.8	9.0	7.0	6.0
Reward Model 1	1.0	0.9	0.8	0.7
Reward Model 2	100	99	80	50
Reward Model 3	9.0	9.8	7.0	6.0

- Spearman Ranking Correlation: **Classification Models!**

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Classification Models

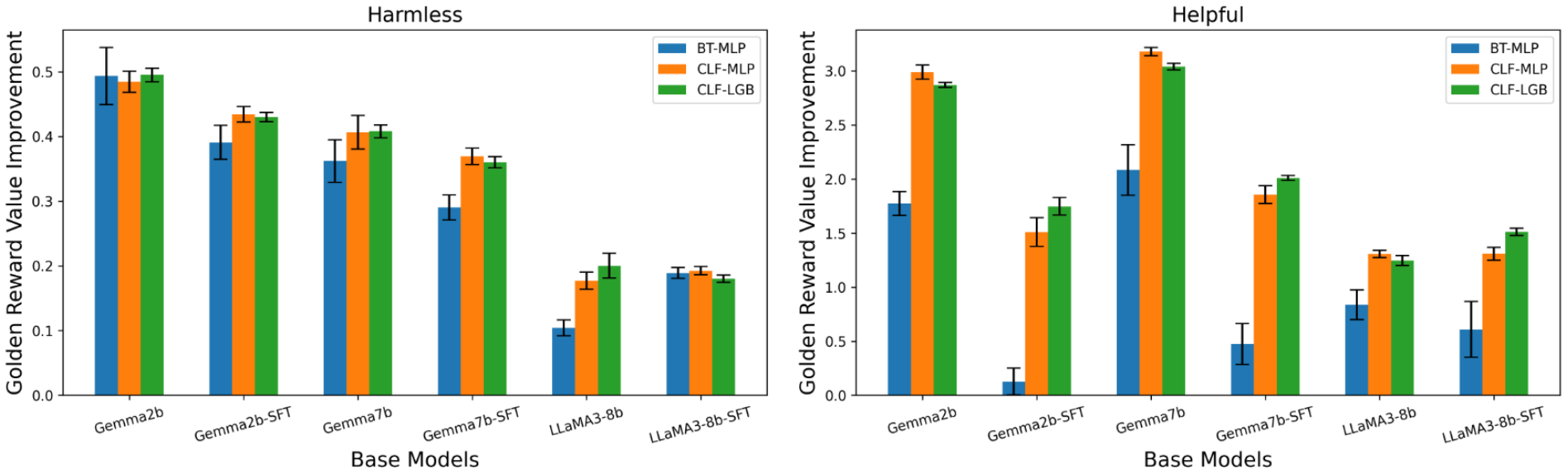
- Classification models are…
  - Flexible

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

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Classification Models

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



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Classification Models

- Classification models are…
  - Flexible
  - Better than BT models (when *labels are noisy*)
  - Robust to annotation noises

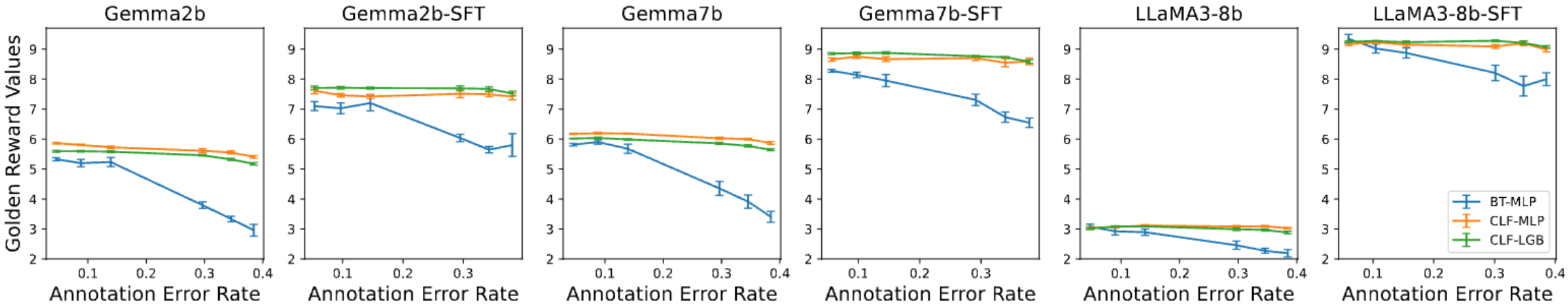


Figure 2: *Changing the annotation quality. Dataset: Harmless, Helpful.*

RM from Preference: Classification Models

- Classification models are…
  - Flexible
  - Better than BT models (when *labels are noisy*)
  - Robust to annotation noises and data scarcity

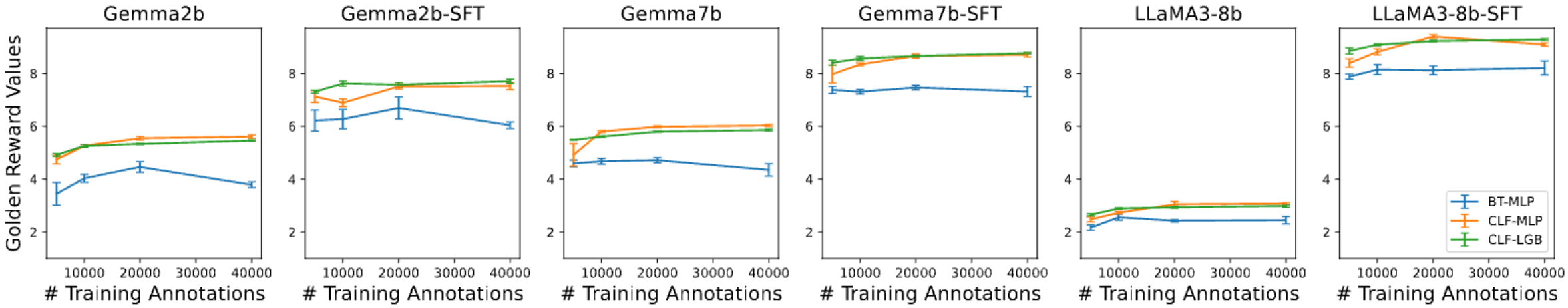


Figure 3: *Changing the annotation quantity. Dataset: Harmless, Helpful.*

RM from Preference: Cross-Prompt Comparisons

- Local Tournaments – Global Tournaments?

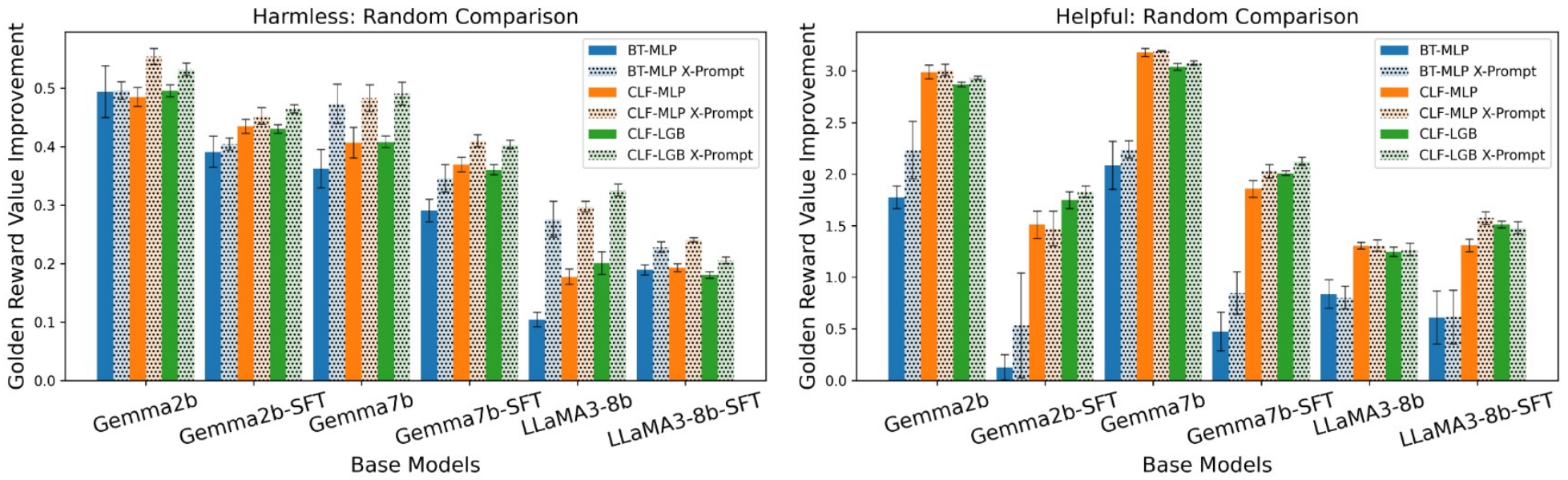


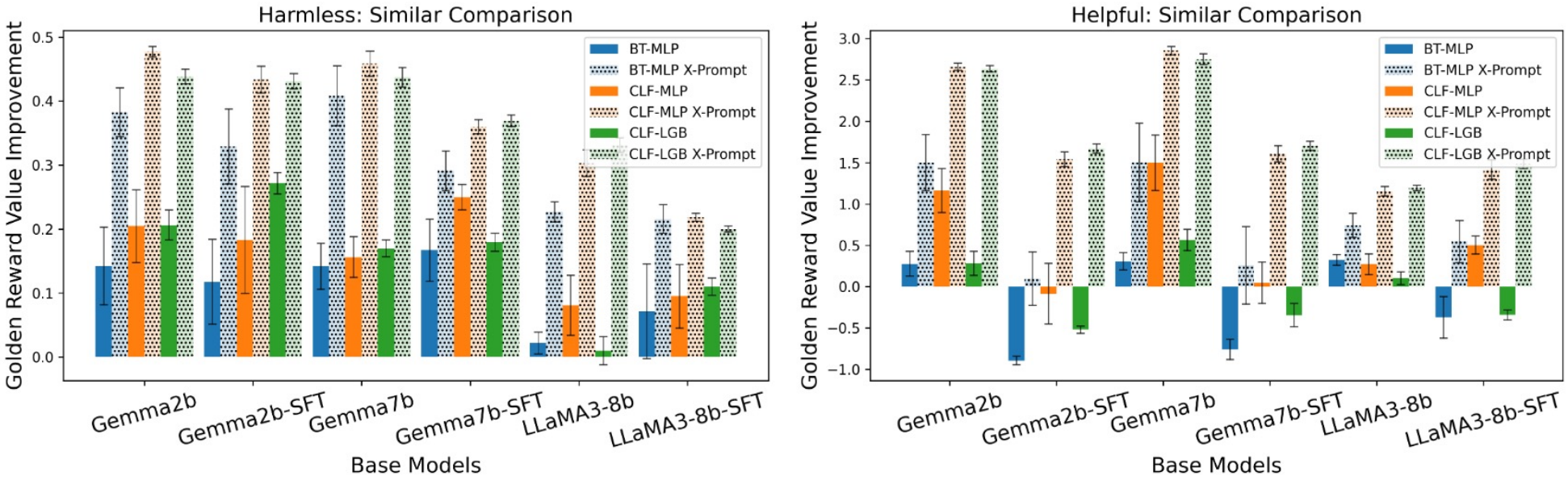
Figure 4: *Results comparing cross-prompt comparison based annotations.* Preference annotations on cross-prompt comparisons outperform same-prompt comparisons.

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Cross-Prompt Comparisons

- When and Why? Generation Diversity Matters

  Low diversity case: clear improvement

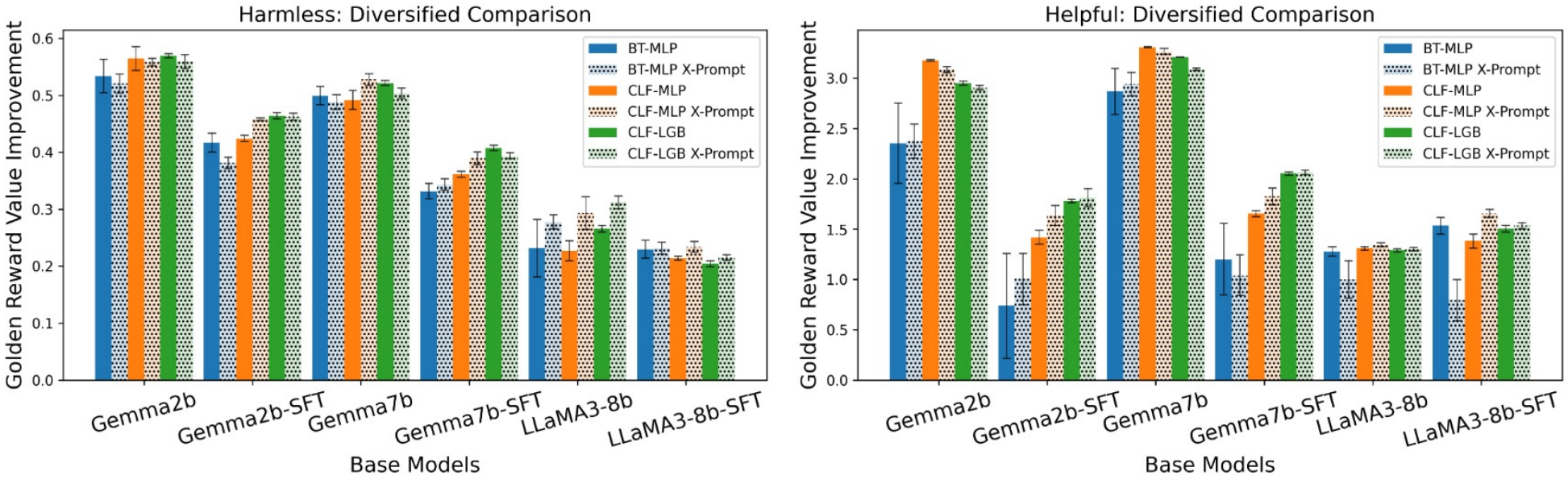


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Cross-Prompt Comparisons

- When and Why? Generation Diversity Matters

  High diversity case: no improvement

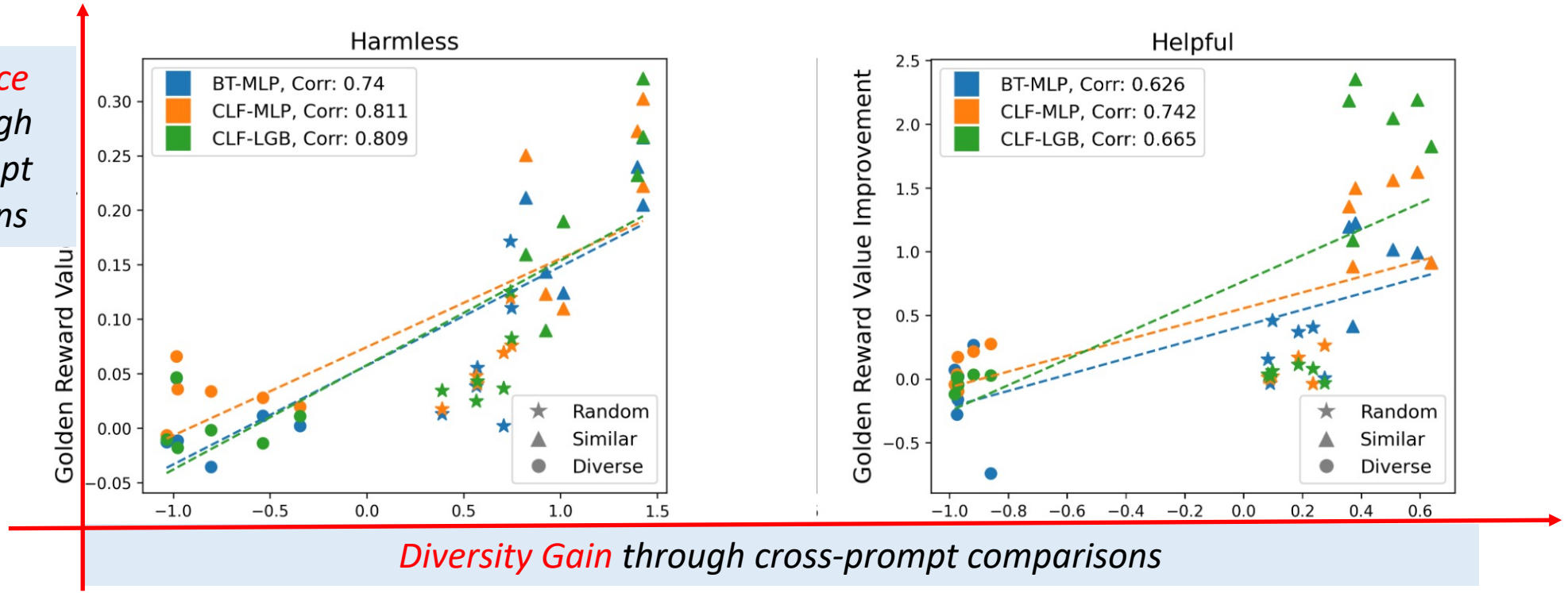


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Cross-Prompt Comparisons

- When and Why? *Comparison* Diversity Matters

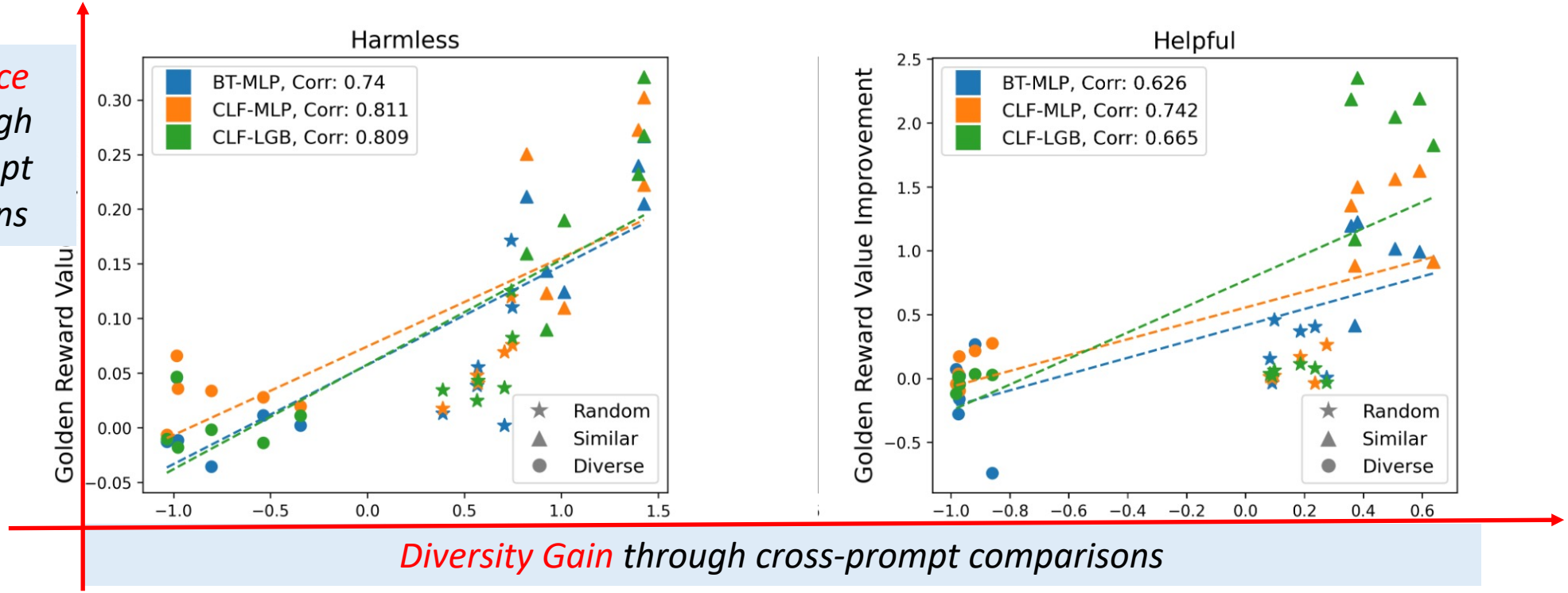
Performance Gain through cross-prompt comparisons



RM from Preference: Cross-Prompt Comparisons

- When and Why? *Comparison* Diversity Matters
  Proof in the paper: Cross-prompt comparisons always improves diversity

Performance Gain through cross-prompt comparisons



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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.
- Use cross-prompt comparisons for diverse comparisons.

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Learning Reward Models from Data

RMs from different data types

RM from Binary (e.g., Math Reasoning)

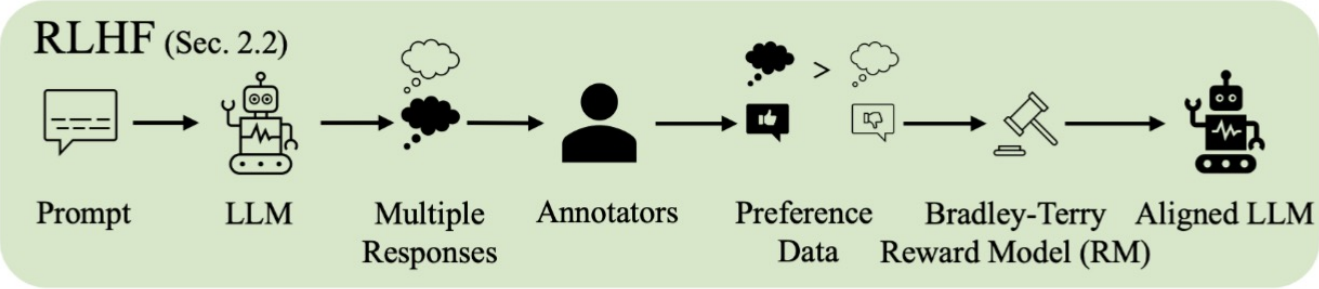
RM from Preference (e.g., Classical RLHF)

RM from Demonstration (e.g., Expert Data)

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

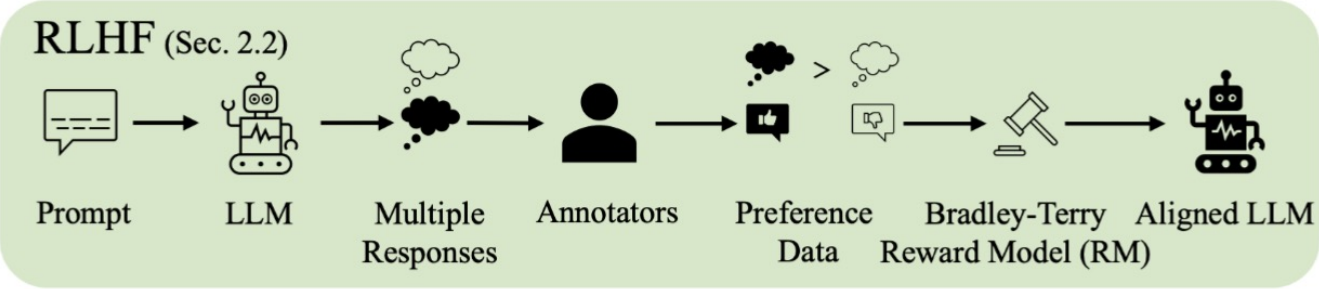
- Common practice of alignment: RLHF



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Common practice of alignment: RLHF

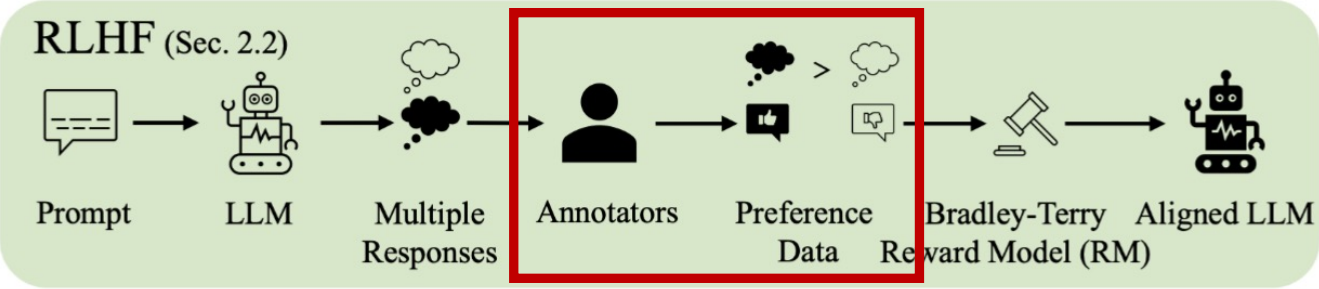


- Why do we need to align LLMs from demonstrations?

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Common practice of alignment: RLHF

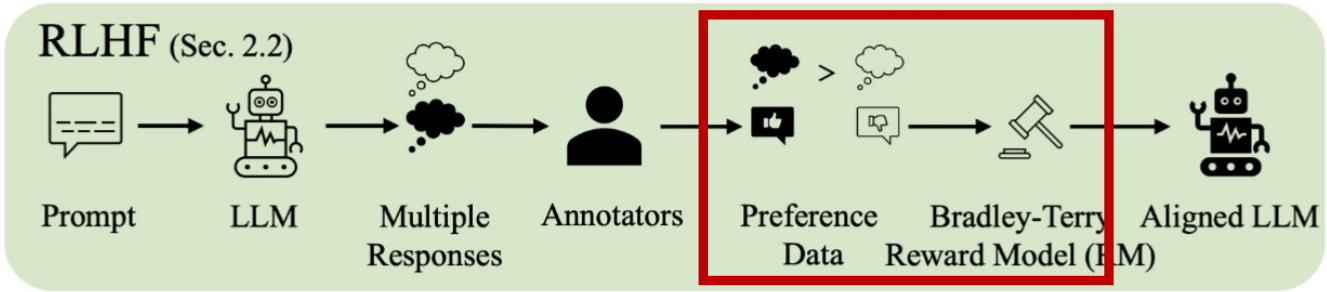


- Why do we need to align LLMs from demonstrations?
  - 1. Preference-based alignment is expensive

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

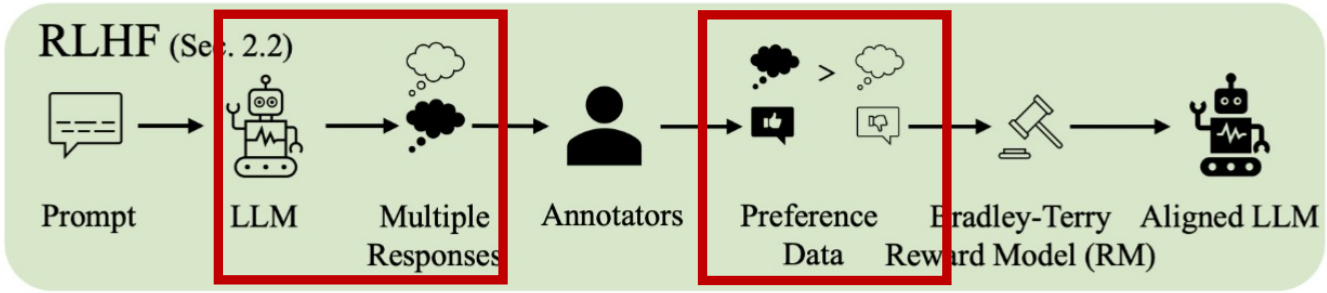
- Common practice of alignment: RLHF



- Why do we need to align LLMs from demonstrations?
  - 1. Preference-based alignment is expensive
  - 2. Assumptions such as Bradley-Terry models are needed

RM from Demonstration: Inverse RL for Alignment

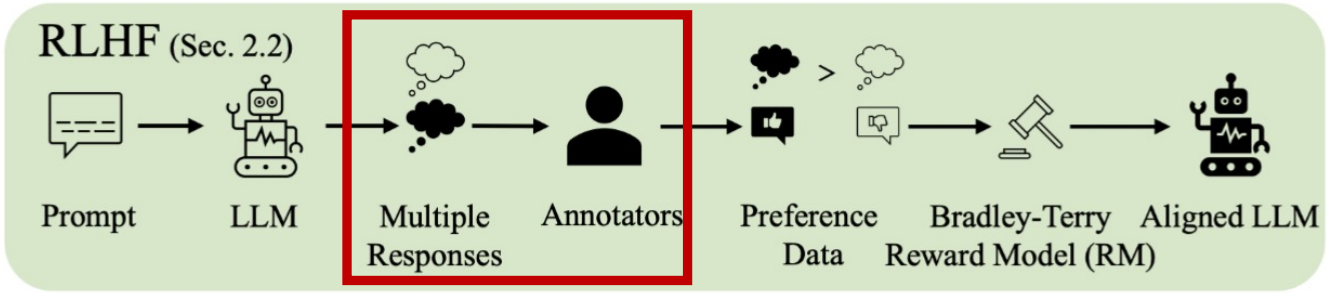
- Common practice of alignment: RLHF



- Why do we need to align LLMs from demonstrations?
  - 1. Preference-based alignment is expensive
  - 2. Assumptions such as Bradley-Terry models are needed
  - 3. Noisy preference annotations

RM from Demonstration: Inverse RL for Alignment

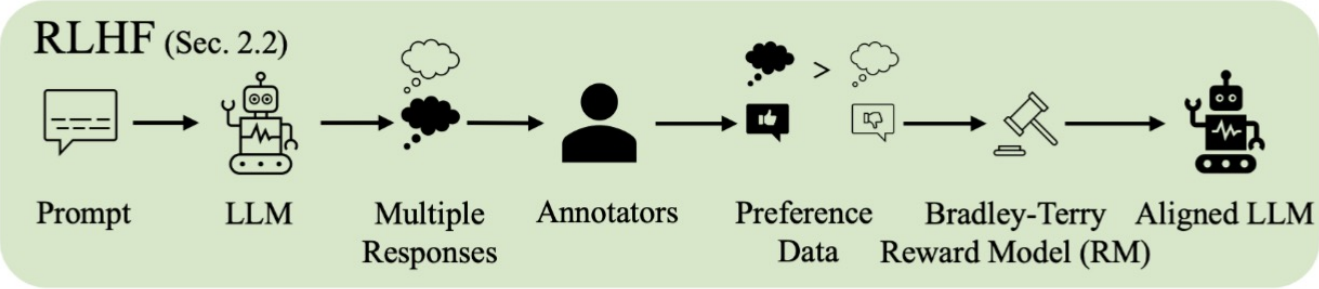
- Common practice of alignment: RLHF



- Why do we need to align LLMs from demonstrations?
  - 1. Preference-based alignment is expensive
  - 2. Assumptions such as Bradley-Terry models are needed
  - 3. Noisy preference annotations
  - 4. Privacy concerns

RM from Demonstration: Inverse RL for Alignment

- Common practice of alignment: RLHF

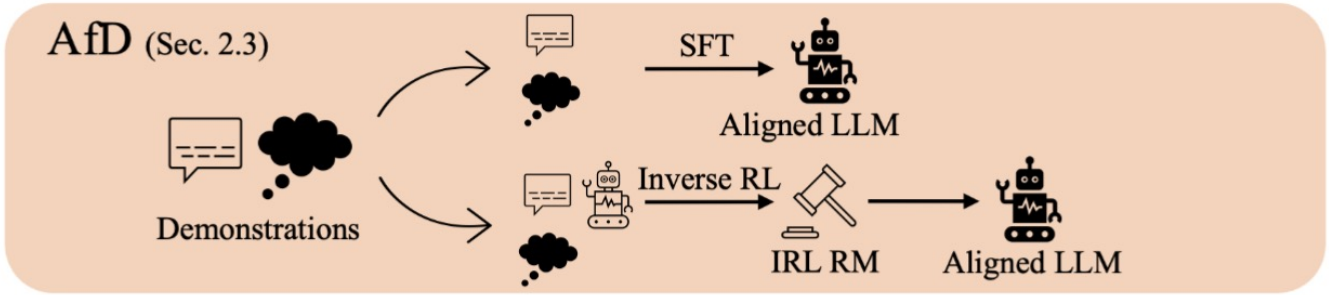


- In applications, we have *high-quality* demonstration data
  - Personalized emails/articles [Shaikh et al., 2024]
  - Medical prescriptions

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Alignment from Demonstrations?



- Vanilla setup in Inverse RL!
  - SFT: Behavior cloning
  - Inverse RL [Sun & van der Schaar, 2024] [Wulfmeier et al., 2024]
  - Imitation Learning [Chen et al., 2024] [Xiao et al., 2024]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- In SPIN [Chen et al., 2024]:

  Lemma 1: Method A can align LLMs using preference data

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- In SPIN [Chen et al., 2024]:

  Lemma 1: Method A can align LLMs using preference data

  Create dataset:

Preferred	---	Demonstration
Dis-preferred	---	Current LLM generated

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- In SPIN [Chen et al., 2024]:

  Lemma 1: Method A can align LLMs using preference data

  Create dataset:

Preferred	---	Demonstration
Dis-preferred	---	Current LLM generated

  Apply *Lemma 1*

  Method A can align LLMs using demonstration data

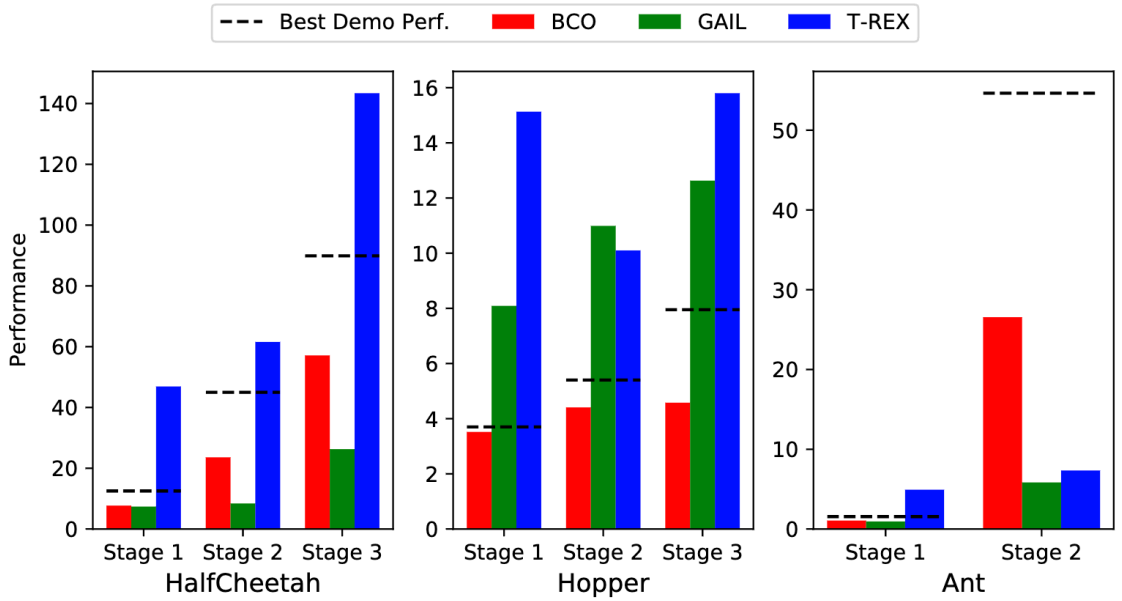
- The objective is to match performance of demonstrations

*Method A in SPIN is DPO [Rafailov et al., 2023]

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Can we do better with explicit reward modeling?
  [Brown et al., 2019]: extrapolation improves performance



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Can we do better with explicit reward modeling?
  [Brown et al., 2019]: extrapolation improves performance

- In [Sun & van der Schaar, 2024] Discriminator as Reward Model:

Create dataset:

Positive	---	Demonstration
Negative	---	Current LLM generated

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Can we do better with explicit reward modeling?
  [Brown et al., 2019]: extrapolation improves performance

- In [Sun & van der Schaar, 2024] Discriminator as Reward Model:

  Create dataset:

  Positive --- ~~Demonstration~~ (*need a fix!*)

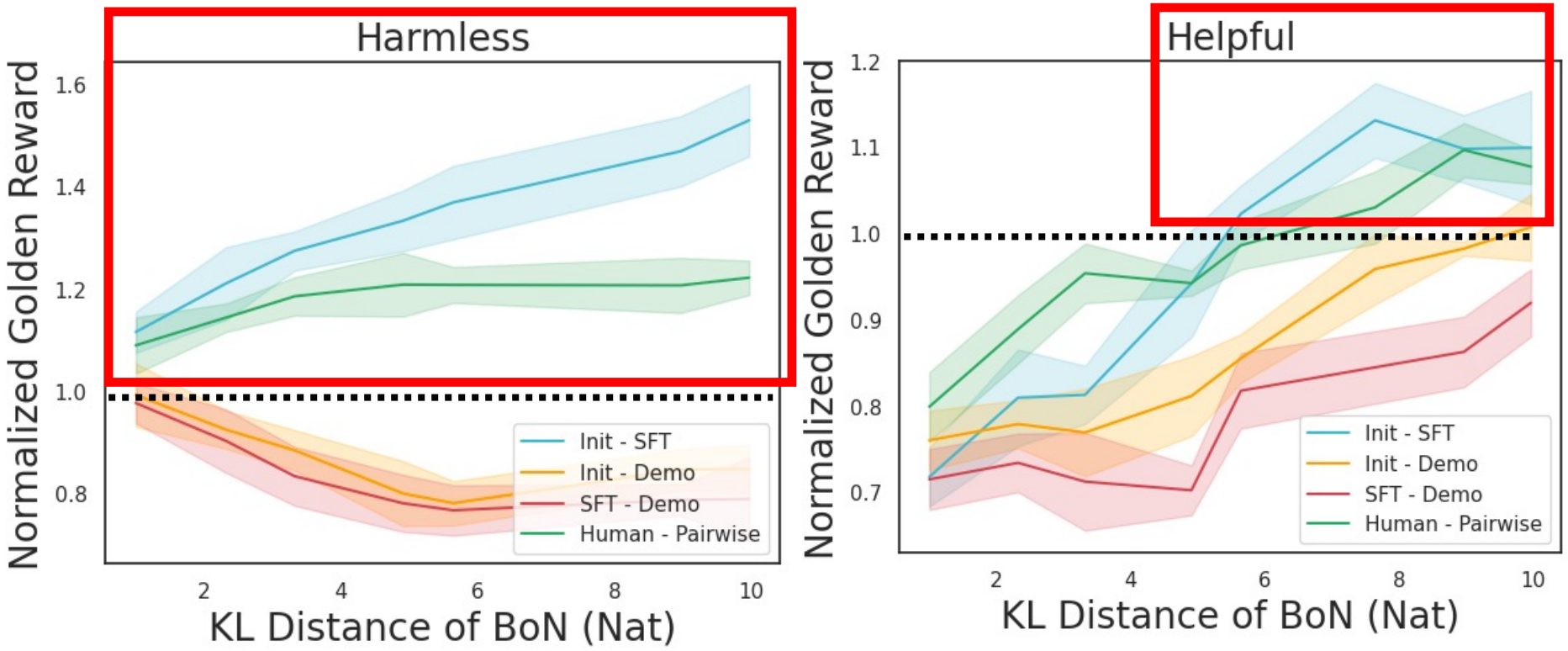
  Negative --- Current LLM generated

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

Super-demonstration performance achieved inside the red boxes.

Dotted lines: Demonstration quality

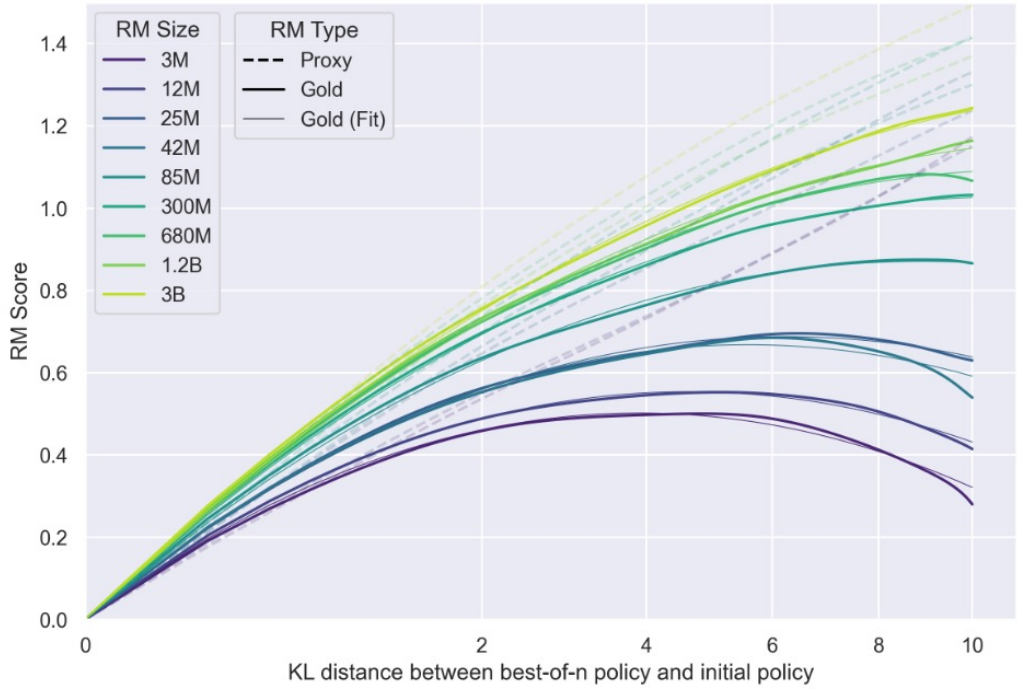


- Super-demonstration performance! (after solving some reward hacking)

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

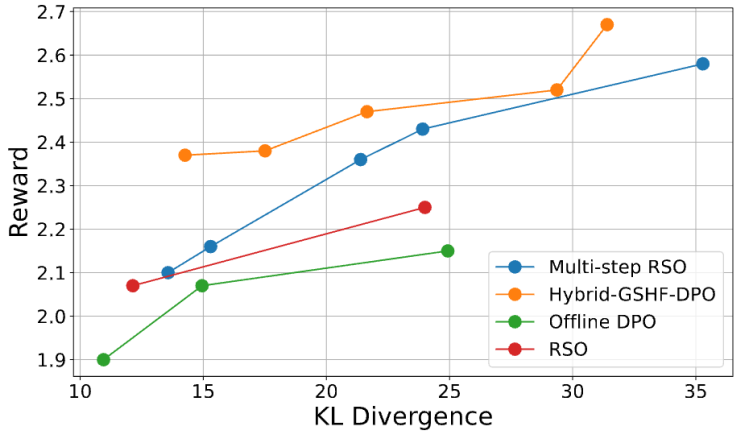
Challenges and Opportunities

- 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

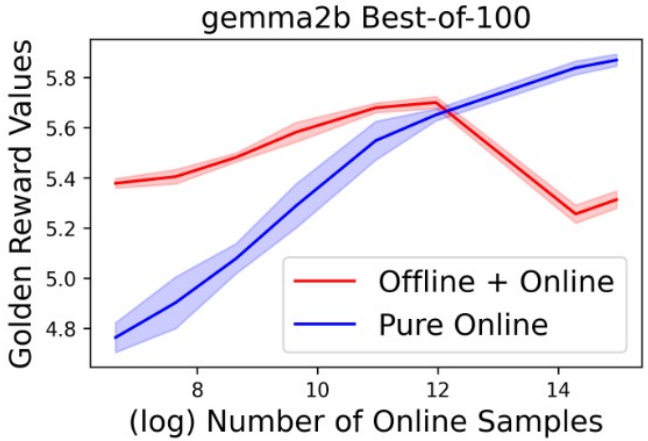
Challenges and Opportunities

- 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]

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 4:

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?

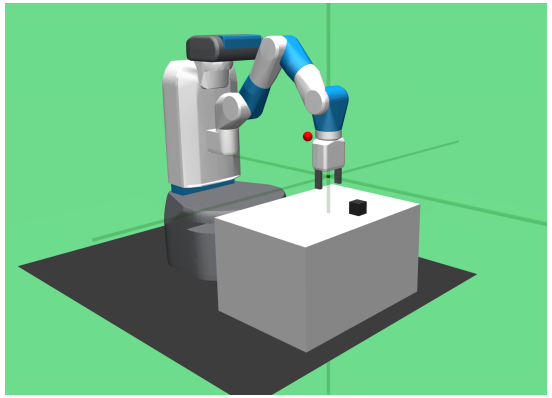
Reward Shaping?

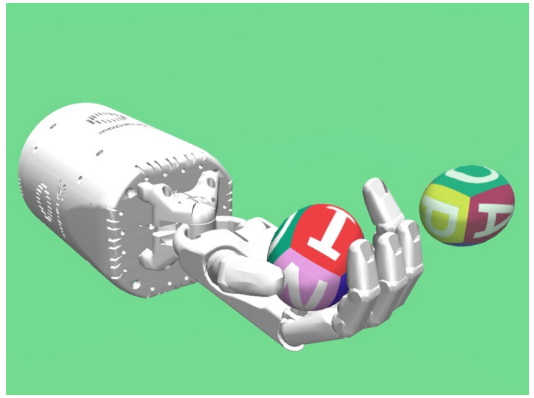
Self-Play?

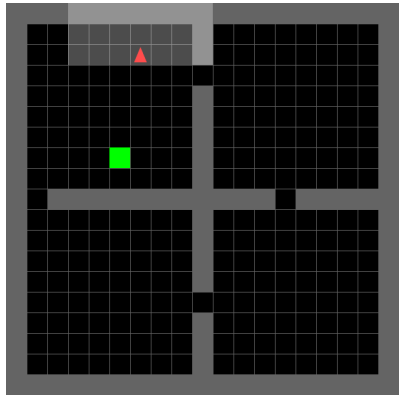
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

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







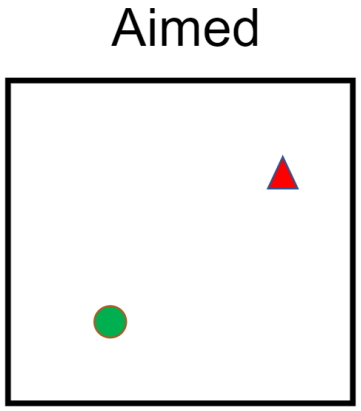
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

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



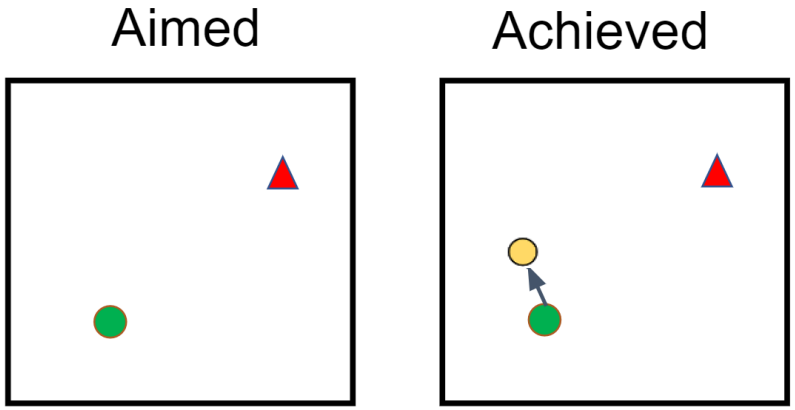
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

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



$a_1,$

$s_1,$

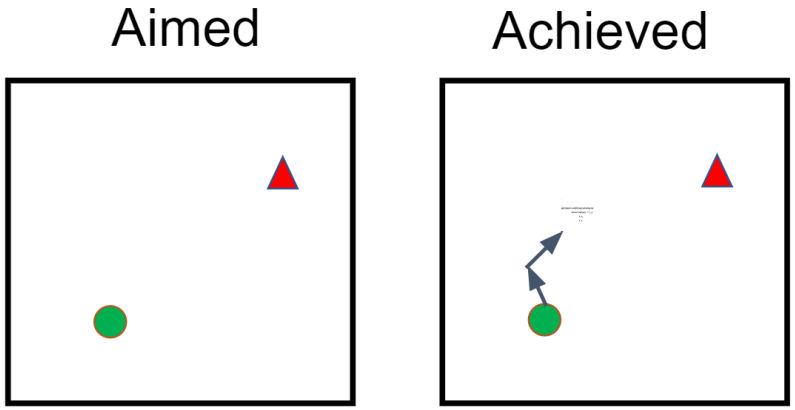
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Hindsight Methods [Andrychowicz et al., 2017]

- How?

- Key insight explained in a simplified supervised learning setup:

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



$a_1, a_2,$

$s_1, s_2,$

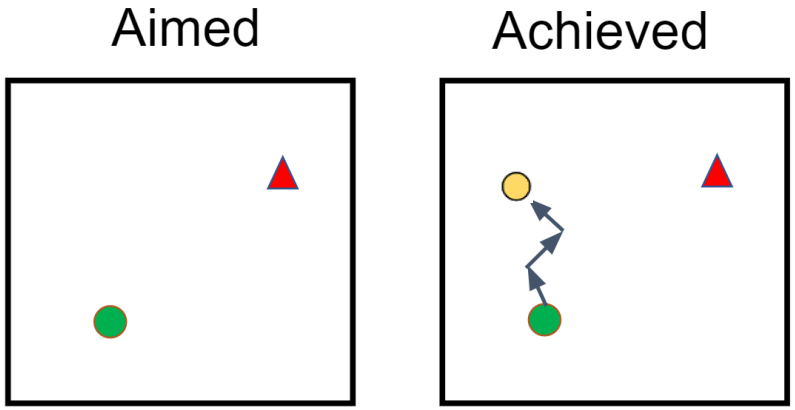
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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

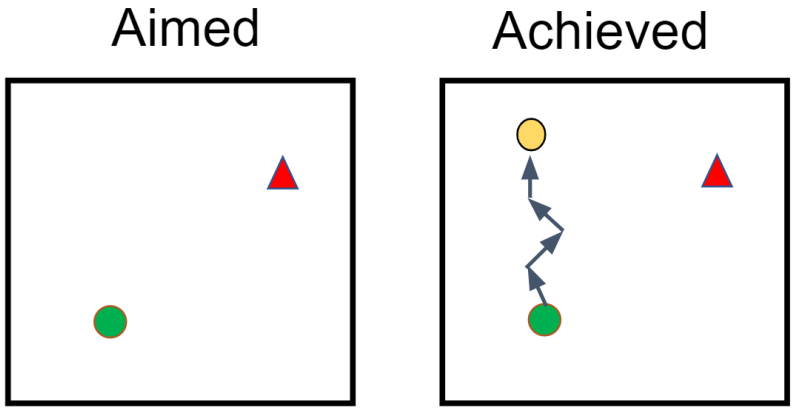
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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

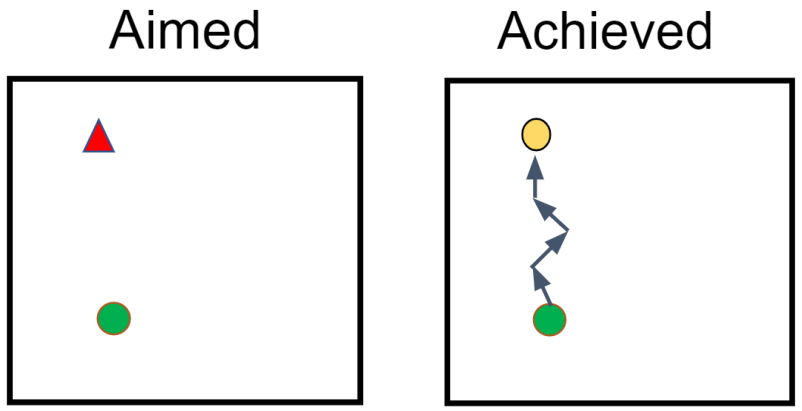
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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

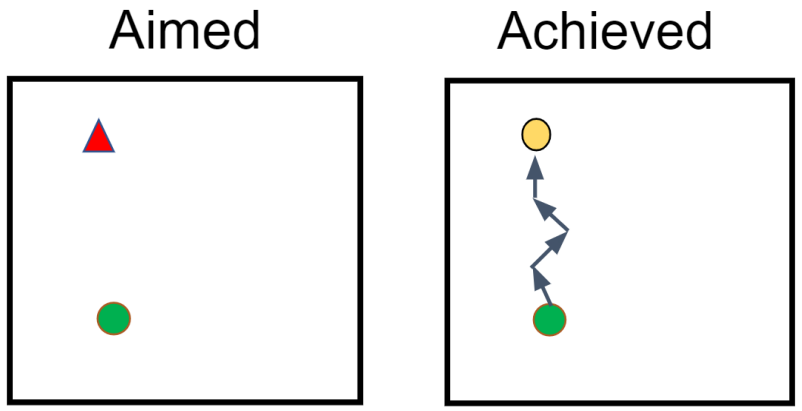
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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 before optimal [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

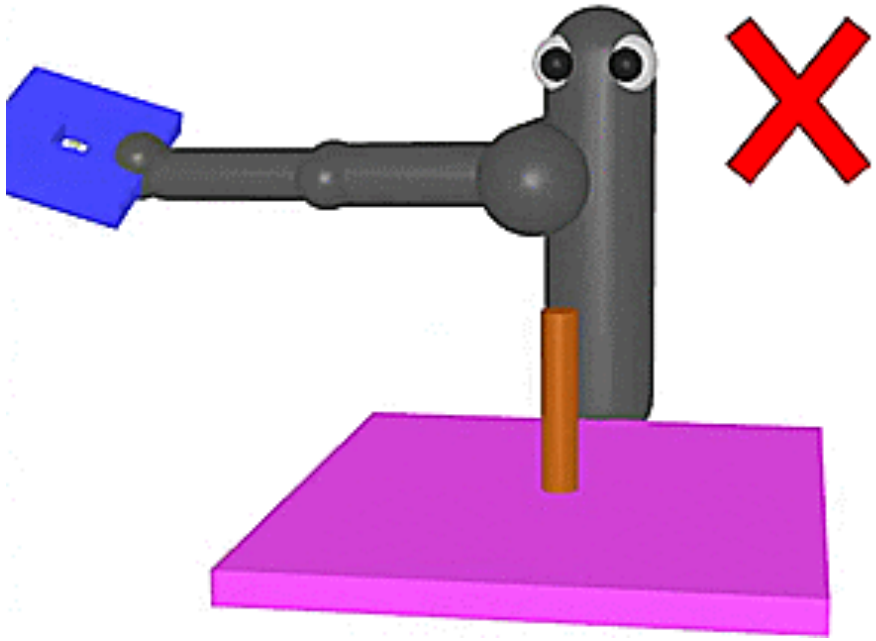
Best Reward: Dense or Sparse?

- 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?

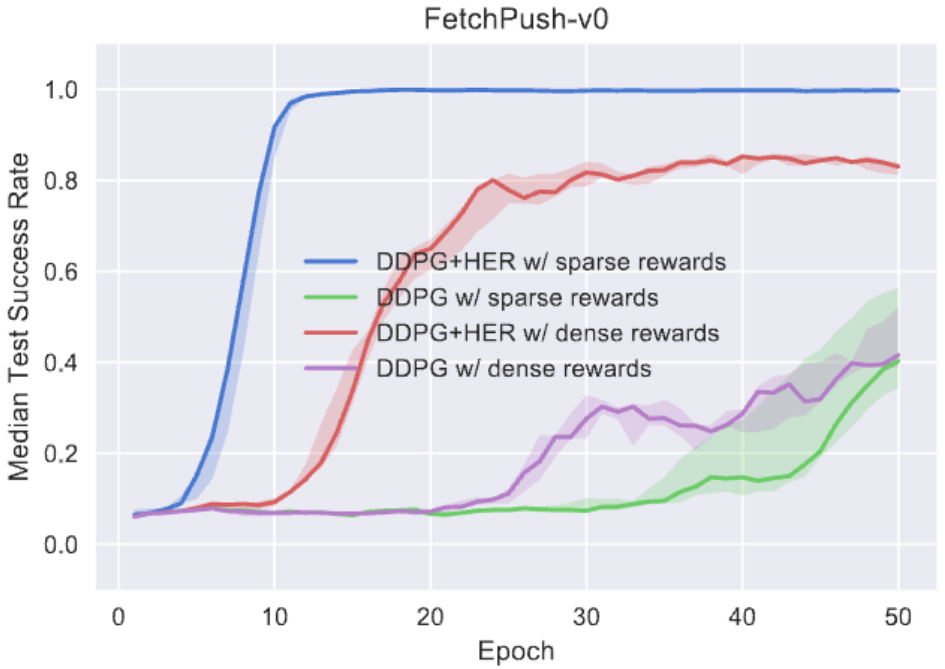
- Learning from sparse reward is challenging --- hard exploration
- Is learning from dense reward always better? --- Not always
  - 1. DeepSeek-R1 [DeepSeek-AI, 2025]
  - 2. Suboptimality [Florensa et al., 2017]



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]



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]
  - 4. Reward hacking [Skalse 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? --- Not always
  - 1. DeepSeek-R1 [DeepSeek-AI, 2025]
  - 2. Suboptimality [Florensa et al., 2017]
  - 3. In multi-goal tasks [Plappert et al., 2018]
  - 4. Reward hacking [Skalse et al., 2022]
- 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]
  - 4. Reward hacking [Skalse et al., 2022]
- Do we need dense reward?
  - Warm-start from demo [Nair et al., 2017]

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]
  - 4. Reward hacking [Skalse et al., 2022]
- Do we need dense reward?
  - No --- Warm-start from demo [Nair et al., 2017]
  - Yes --- AlphaGo

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?

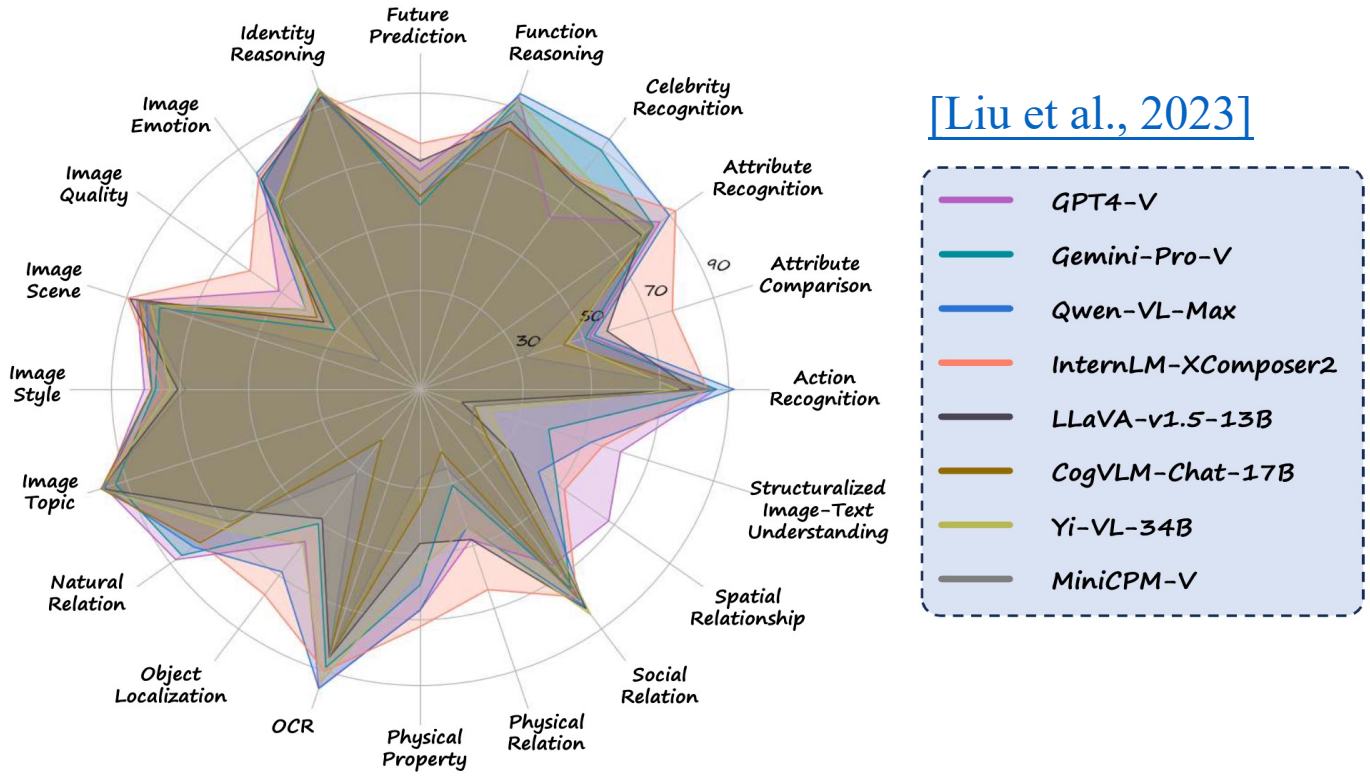
Task	Action Space	State Space	Reward Signal	Method	Transition
Go	Disc. ~ 1e2	Disc. ~1e100	Sparse	MCTS, Self-Play	Known
Reasoning	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Sparse**	GRPO	**Known**
Alignment	**Disc. ~ 1e6**	**Vocab.^Token_n**	**Noisy Preference**	PPO, DPO, REINFORCE	**Known**

- 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