Inverse RL Meets LLMs:

RL for better Prompting, Fine-Tuning, and Inference-Time Optimization

Hao Sun

Oct 2024

van_der_Schaar \ LAB

vanderschaar-lab.com

UNIVERSITY OF CAMBRIDGE

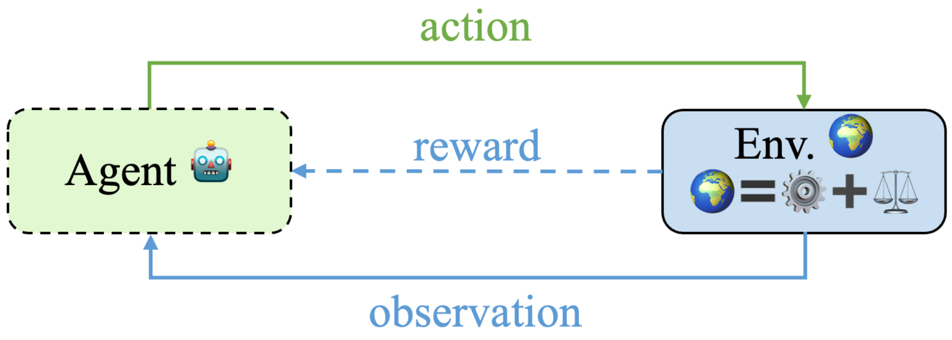
hs789@cam.ac.uk

@HolarisSun

Content

Concepts in (Inverse) RL

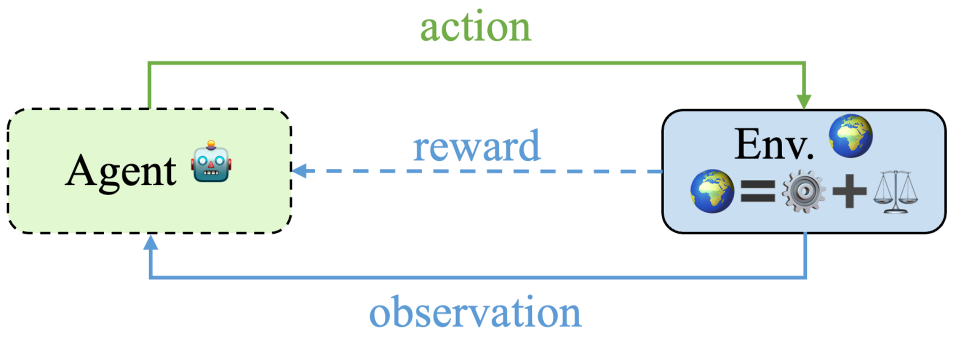
- Environment = Dynamics + Reward



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

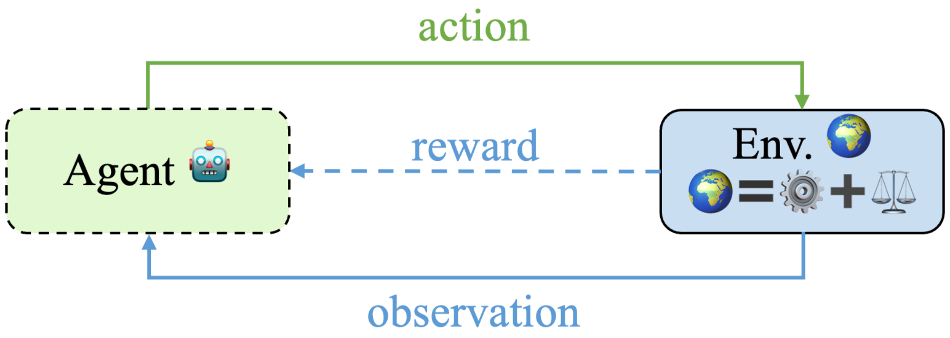
Concepts in (Inverse) RL

- Environment = Dynamics + Reward
- Learning from trial and error: *execution* & *evaluation* are expensive



Concepts in (Inverse) RL

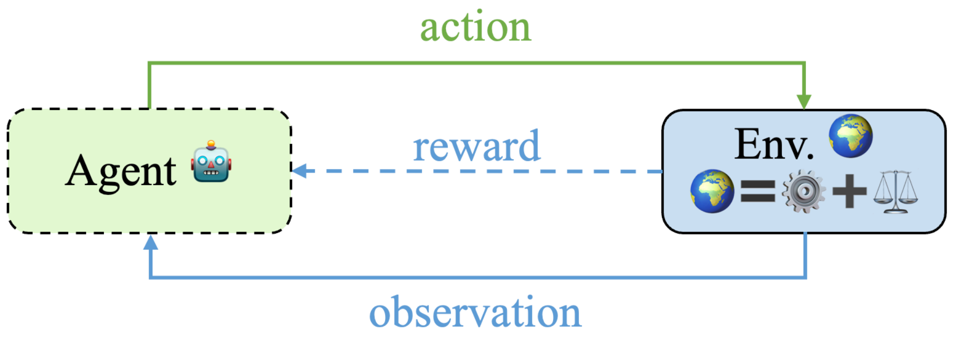
- Environment = Dynamics + Reward
- Learning from trial and error: *execution & evaluation* are expensive
- Learning by imitating is easier: *Behavior Clone*

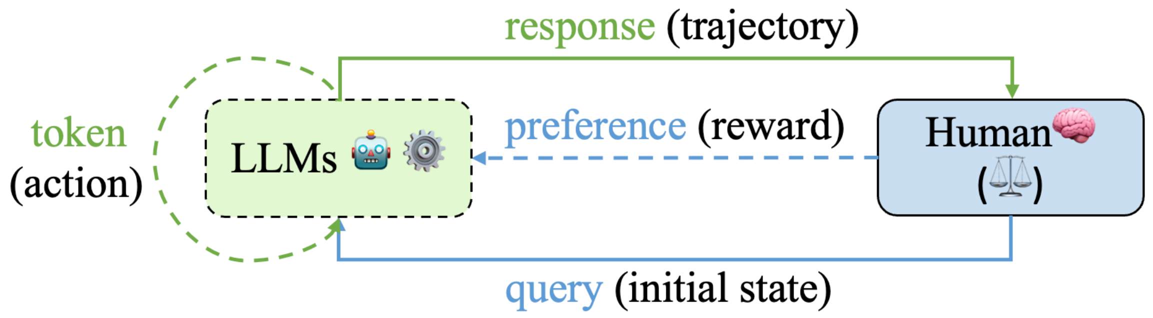


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Concepts in (Inverse) RL

- Environment = Dynamics + Reward
- Learning from trial and error: *execution & evaluation* are expensive
- Learning by imitating is easier: *Behavior Clone*
- LLM is a special case --- only reward is expensive



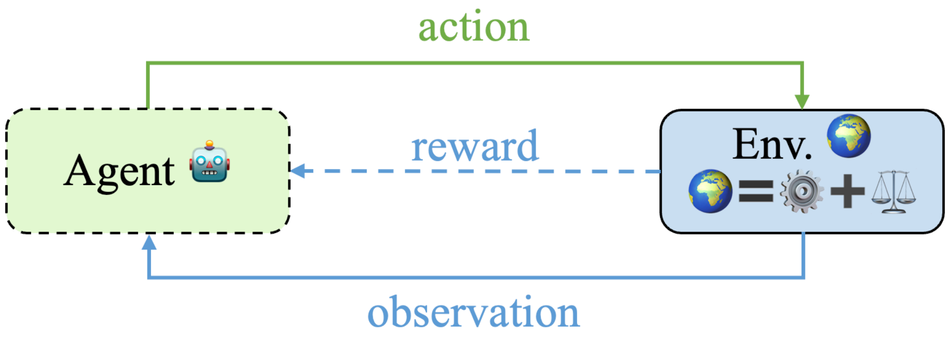


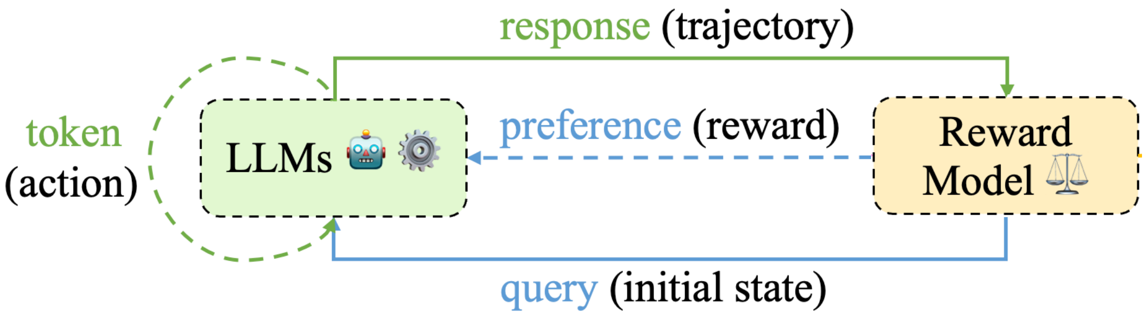
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Concepts in (Inverse) RL

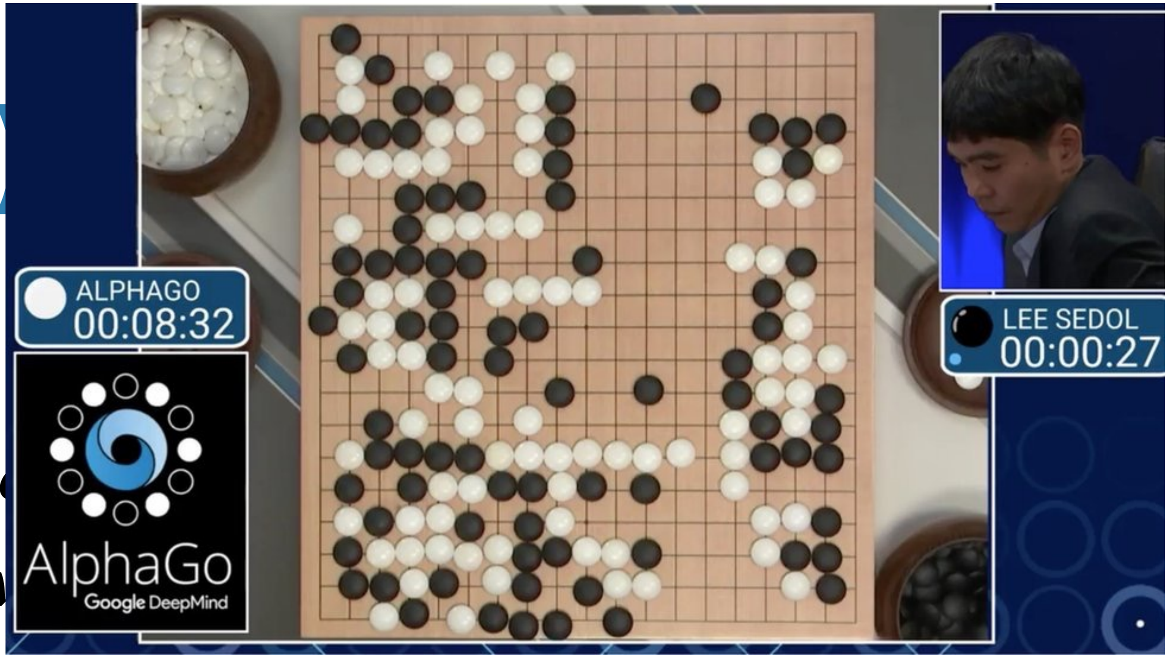
- Environment = Dynamics + Reward
- Learning from trial and error: *execution & evaluation* are expensive
- Learning by imitating is easier: *Behavior Clone*
- LLM is a special case --- only reward is expensive

Reward Modeling is essential!



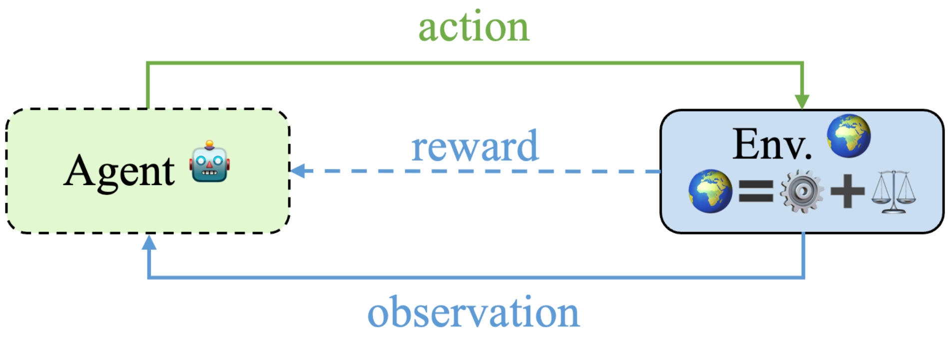


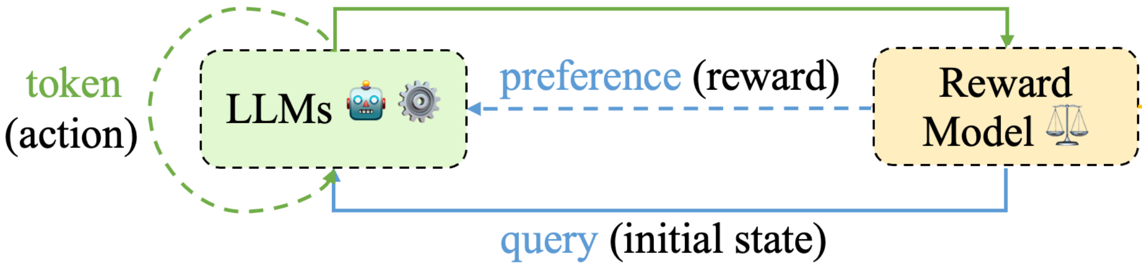
Concepts in (Inverse)



- Environment = Dynamics + Reward
- Learning from trial and error: *executi*
- Learning by imitating is easier: *Behav*
- LLM is a special case --- only reward is expensive

Reward Modeling is essential!





hs789@cam.ac.uk
UNIVERSITY OF
CAMBRIDGE

LLMs: Language Imitators

- How are LLMs trained?
  - Pre-training phase: supervised-learning
  - Post-training phase: SFT, RLHF …

hs789@cam.ac.uk
UNIVERSITY OF
CAMBRIDGE

LLMs: Language Imitators

- How are LLMs trained?
  - Pre-training phase: supervised-learning
  - Post-training phase: SFT, RLHF …

behavior cloning

--- from the perspective of Inverse RL

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLMs: Language Imitators

- How are LLMs trained?
  - Pre-training phase: supervised-learning
  - Post-training phase: SFT, RLHF …

--- Imitating Natural Language

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

LLMs: Language Imitators

- How are LLMs trained?
  - Pre-training phase: supervised-learning
  - Post-training phase: SFT, RLHF …

--- **Imitating Natural Language**

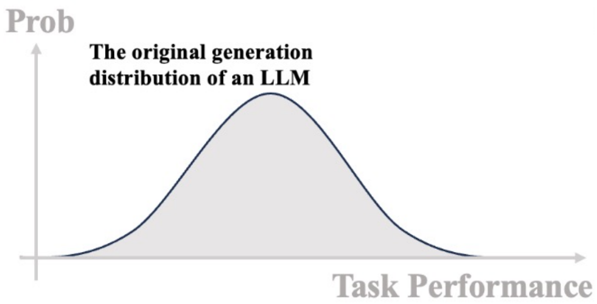
- What are LLMs?
  - LLMs can provide approximate knowledge [1]
  - But they are MERELY experts for ANYTHING.

We need **Reward Models**

[1] Kambhampati, Subbarao, et al. "LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks." *arXiv preprint arXiv:2402.01817* (2024).

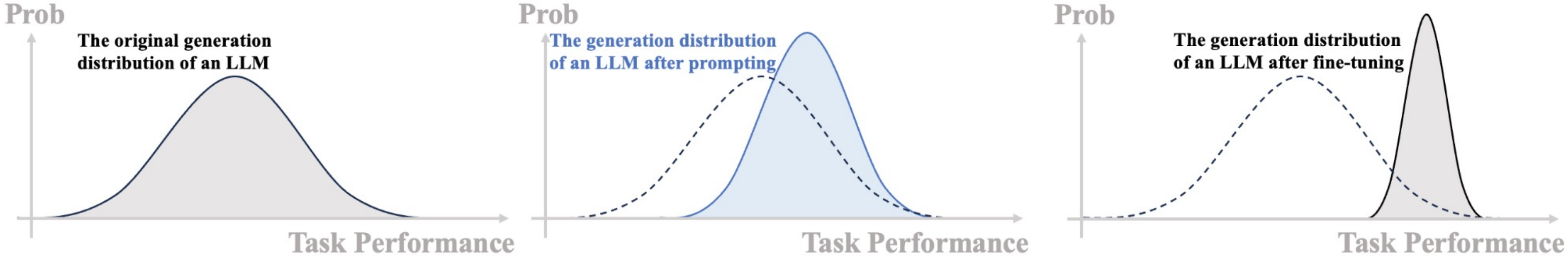
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Optimizing LLM Usages with RMs



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

Optimizing LLM Usages with RMs



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

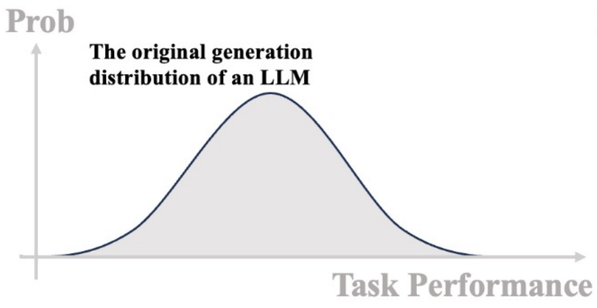
(2) Prompting *Can* Improve Performance by shifting the generation

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

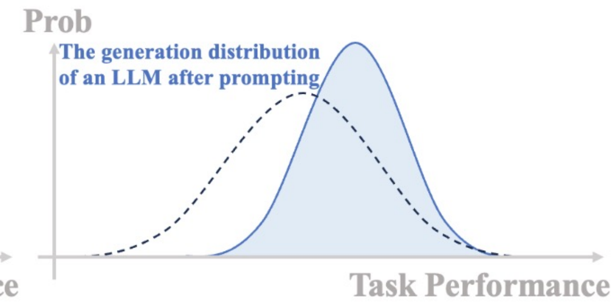
hs789@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

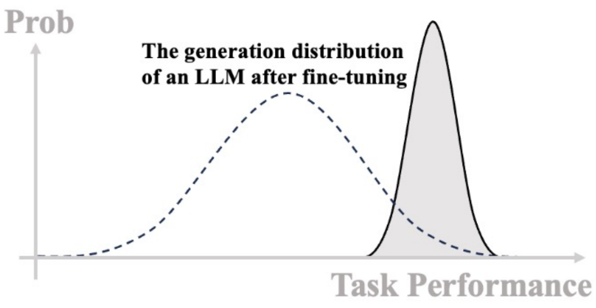
Optimizing LLM Usages with RMs



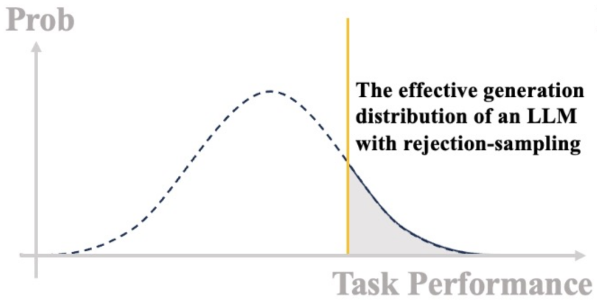
(1) LLM *Can* do Any Task as a *Universal Sampler*



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



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

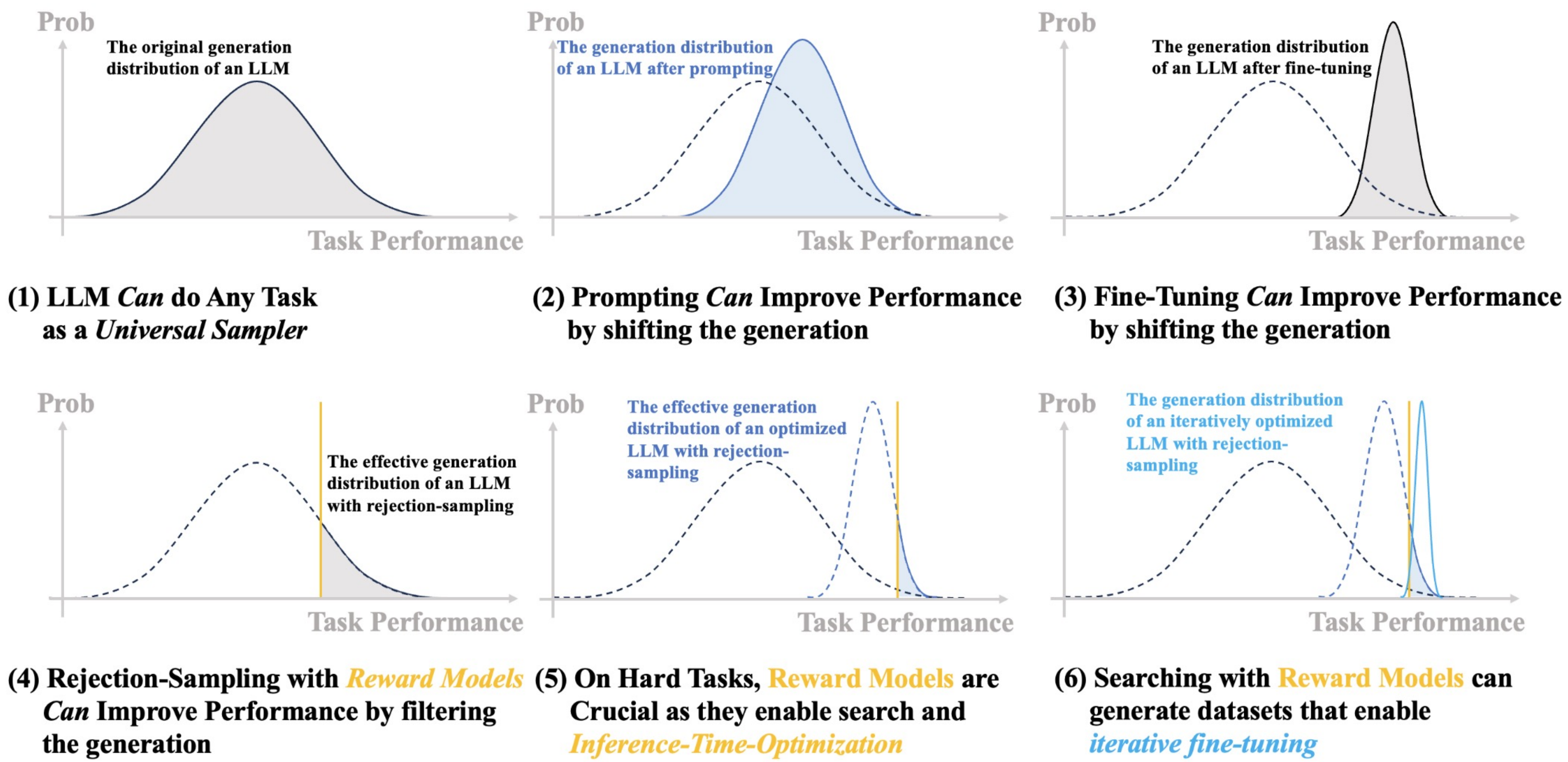


(4) Rejection-Sampling with *Reward Models* *Can* Improve Performance by filtering the generation

Reward Models
Enable Inference-time Optimization

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Optimizing LLM Usages with RMs



Reward Models
Enable Inference-time Optimization

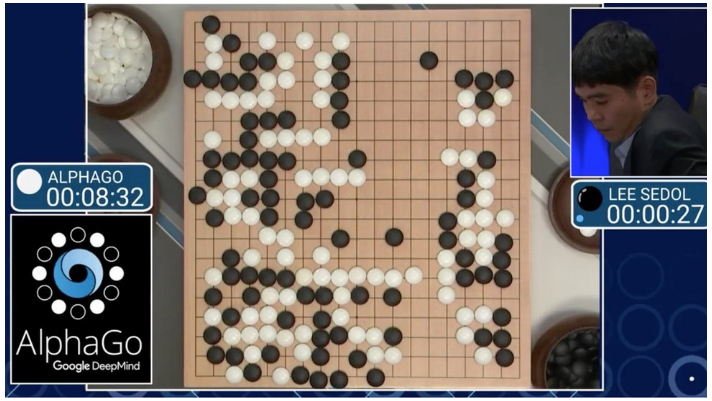
Takeaways:

- Reward models --- Foundation of LLM Optimization

hs789@cam.ac.uk
UNIVERSITY OF
CAMBRIDGE

Takeaways:

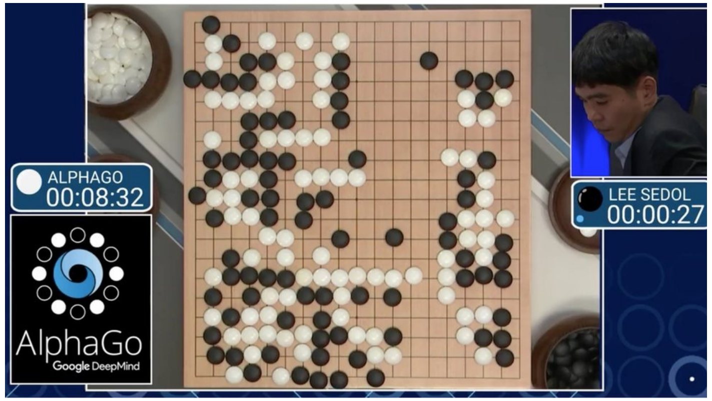
- Reward models --- Foundation of LLM Optimization
  - From an RL Perspective,
    RM is the only missing part of an “RL-solvable task”

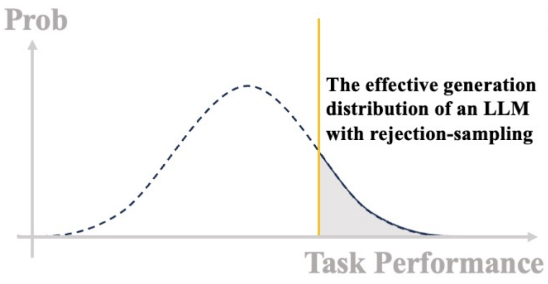


hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

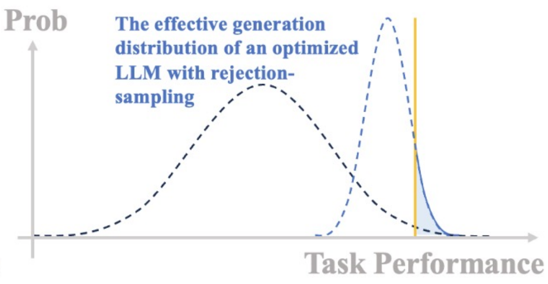
Takeaways:

- Reward models --- Foundation of LLM Optimization
  - From an RL Perspective,
    RM is the only missing part of an "RL-solvable task"
  - From an LLM Perspective,
    RM enables Inference-time Optimization

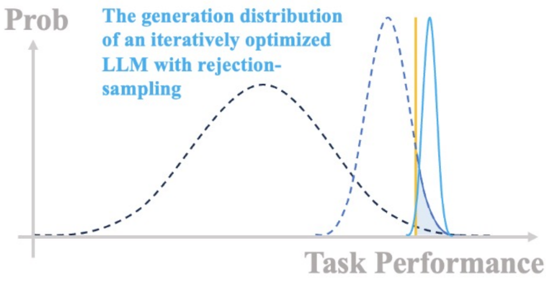




(4) Rejection-Sampling with *Reward Models* *Can* Improve Performance by filtering the generation



(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*

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL x LLM: The Inverse and Forward

Inverse direction

Forward direction

→ **Reward Models** →

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RL x LLM: The Inverse and Forward

Inverse direction

Forward direction

Binary (Reasoning)

[Prompt-OIRL]

Preference Data

[RMBeyondBT]

Demonstration

[InverseRLignment]

→ **Reward Models** →

hs789@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

RL x LLM: The Inverse and Forward

Inverse direction

Binary (Reasoning)

[Prompt-OIRL]

Preference Data

[RMBeyondBT]

Demonstration

[InverseRLignment]

→ **Reward Models** →

Forward direction

When are RMs Useful?

[DataCOPE]

Process Reward

[DenseReward]

Reasoning with MCTS

[RATP]

Key of ChatGPT: zero-shot CoT, RLHF, On-policy RMs

Inverse direction		Forward direction
Binary (Reasoning)		When are RMs Useful?
Preference Data	→ Reward Models →	Process Reward
Demonstration		Reasoning with MCTS

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Key of o1: Process Reward, Search-based generation, RMs

Inverse direction

Binary (Reasoning)

Preference Data

Demonstration

Reward Models

Forward direction

When are RMs Useful?

Process Reward

Reasoning with MCTS

Content

- Preliminaries
  - Key Concepts in RL and Inverse RL
  - Key Concepts in LLMs
- Inverse RL Meets LLMs
  - What Makes ChatGPT Great
  - What Makes o1 Better
- Building Reward Models from Data (Inverse)
  - Binary data: Offline Inverse RL for Mathematical Reasoning
  - Preference data: Foundations of Preference-based Reward Modeling
  - Demonstration data: Inverse RL for Alignment from Demonstrations

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

- Prompt engineering is useful, but empirical…

(User Directly Asks the Question)	(User Uses Multi Agent Debate Prompting)
a^4 = 1, what is a?	a^4 = 1, what is a? Two experts are debating on the answer:
(GPT-4 gives the *correct* answer)	(GPT-4 gives a *wrong* 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 $-i$.	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.

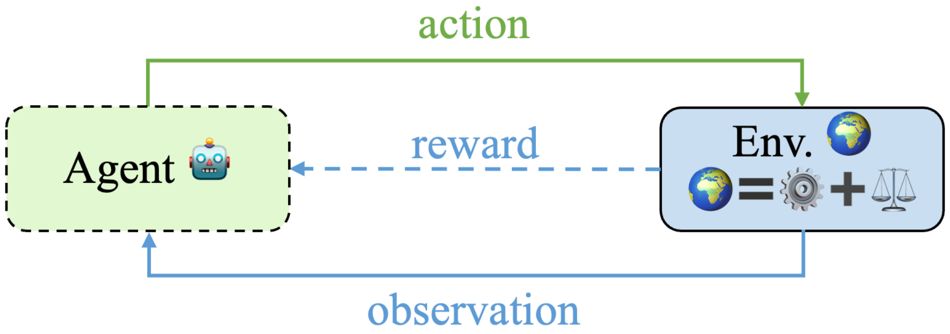
- Automatic prompt engineering is expensive…
  - Prompt-Dependent
  - Huge vocabulary space

hs789@cam.ac.uk

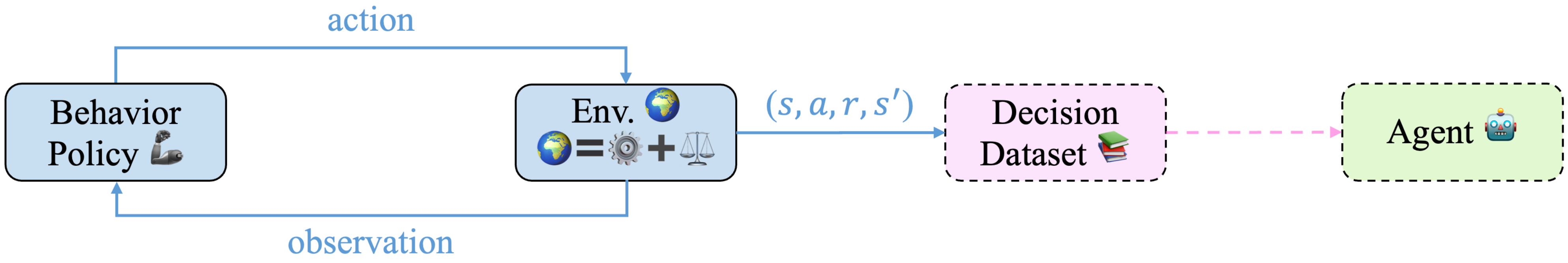
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

- Learning from demonstrations can be more efficient than from scratch
- RL Recap: Learning from interactions



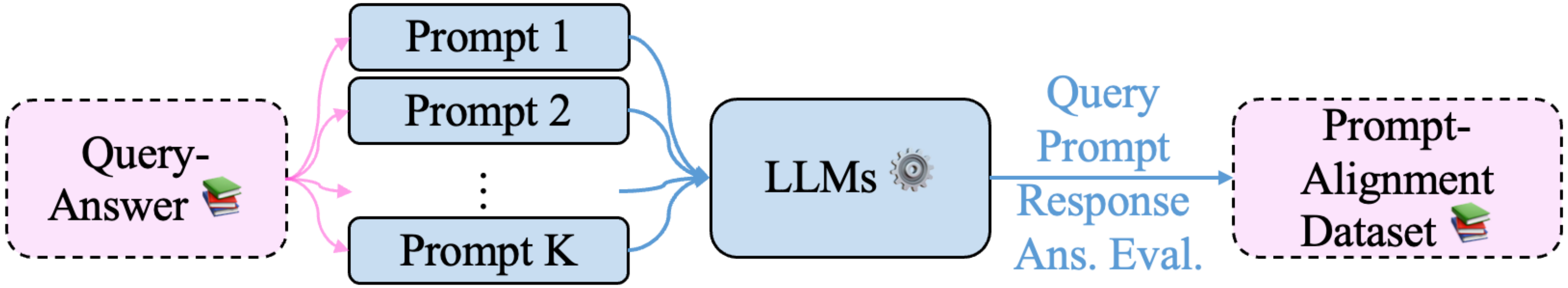
- Prompt Optimization as Inverse RL: learning from demonstrations



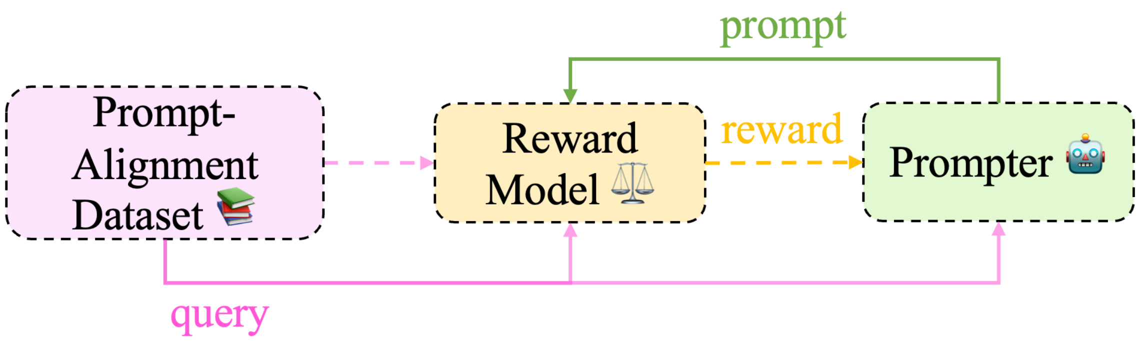
hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Binary Data: IRL for *Prompt Optimization*

- Offline Inverse RL for Prompt Optimization
- Existence of demonstration dataset:

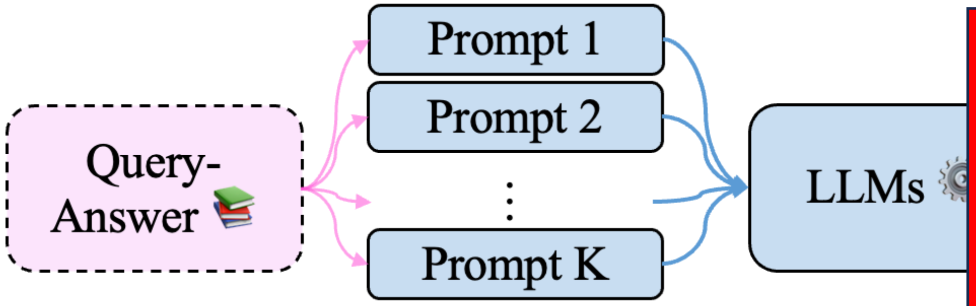


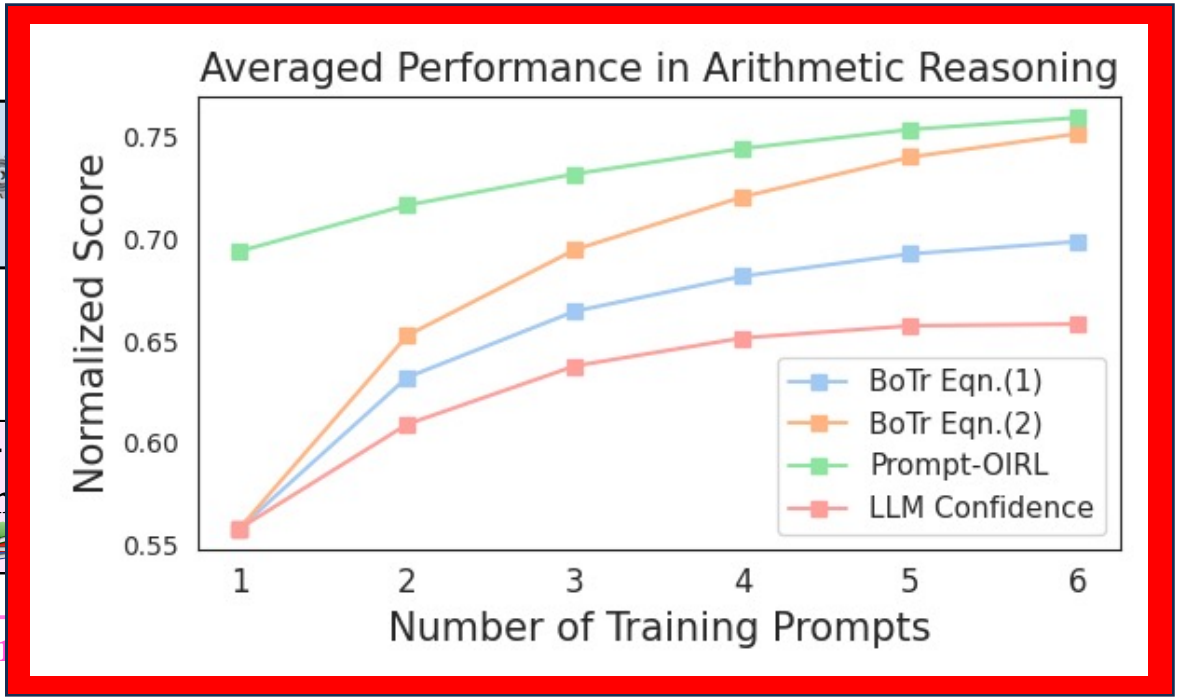
- Inverse RL:
  - Reward Modeling
  - Prompt Optimization



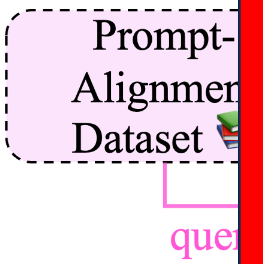
RM from Binary Data: IRL for *Prompt Optimization*

- Offline Inverse RL for Prompt Optimization
- Existence of demonstration dataset: **Math Reasoning Ability**





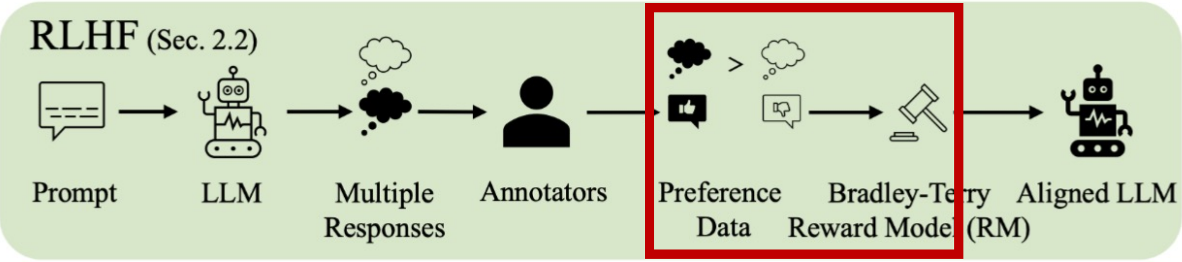
- Inverse RL:
  - Reward Modeling
  - Prompt Optimization



hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

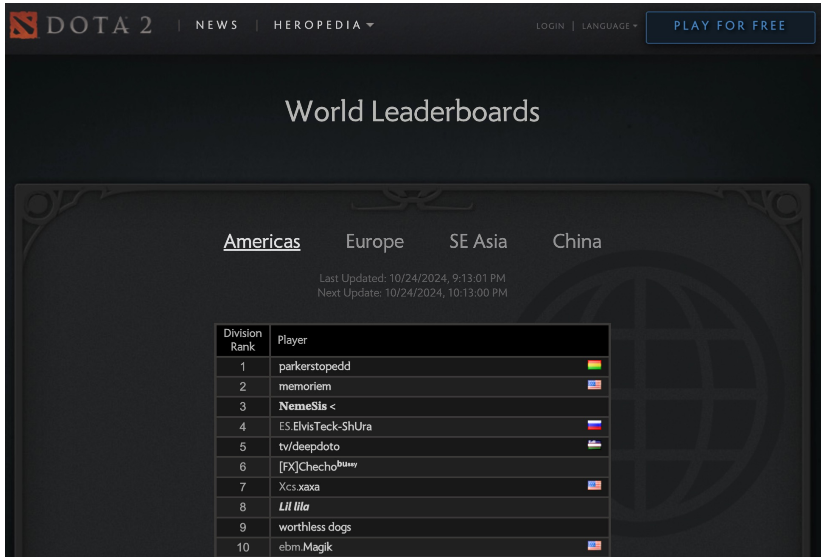
- In chat tasks, online preference data can be available
- How to build reward model?
  - RLHF: "use the Bradley-Terry Model"
  - But why?



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

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 26,000+ games



Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization
1	1	ChatGPT-4o-latest (2024-09-03)	1340	+4/-5	31927	OpenAI
1	1	o1-preview	1337	+4/-5	19924	OpenAI
3	5	o1-mini	1309	+5/-4	21425	OpenAI
3	3	Gemini-1.5-Pro-002	1303	+5/-5	13957	Google
4	3	Gemini-1.5-Pro-Exp-0827	1299	+4/-3	32393	Google
6	8	Grok-2-08-13	1290	+4/-4	39193	xAI
6	11	Yi-Lightning	1286	+4/-4	18864	01 AI
6	3	GPT-4o-2024-05-13	1285	+3/-2	101733	OpenAI
9	14	GLM-4-Plus	1275	+5/-4	18695	Zhipu AI

van_der_Schaar \ LAB

vanderschaar-lab.com

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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 26,000+ games

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

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

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 26,000+ games

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

- In RLHF,
  - we have m prompts, 2m responses --- 2m players
  - Each pair only “compete” once --- $m \ll 2m\log(2m)$ comparisons
  - ⚠️ We have more than 2m parameters to estimate --- ⚠️⚠️⚠️ we need predictions!

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

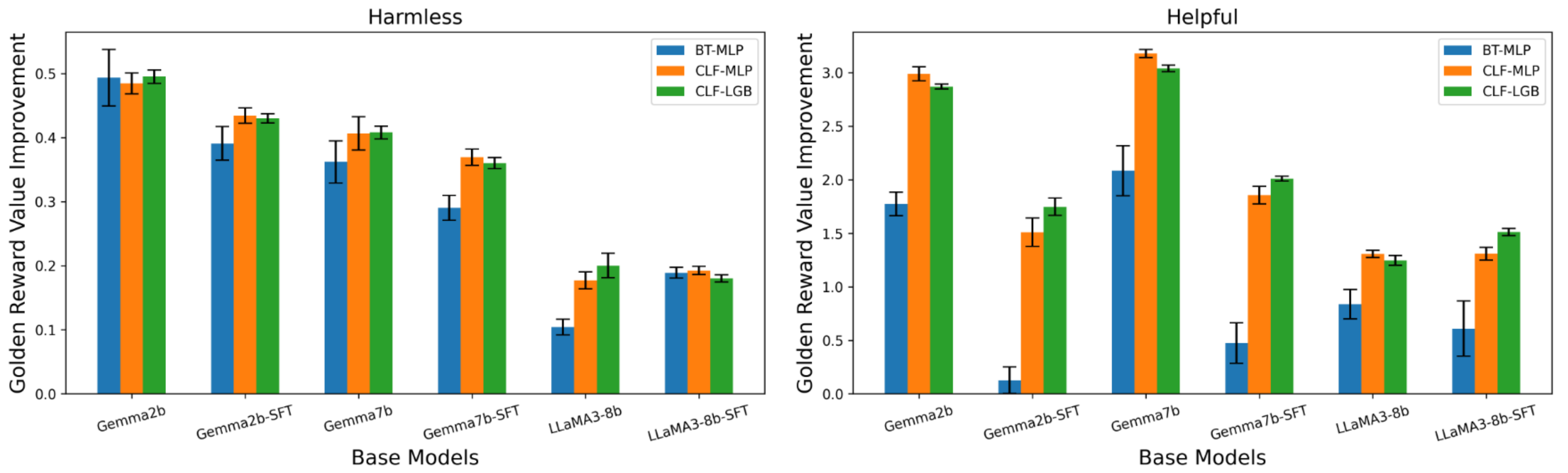
- Why does BT model work?
  - We are working on the embedding space
  - Generalizable…
  - Theoretical justification is in the paper

- Is BT model necessary?
  - Rethinking the objective of BT model: precisely predicting win rates
  - Is it necessary in RLHF?
  - NO, we only need order consistency --- order of prediction is aligned with data
  - Binary classification ☺

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Preference: Back to Ranking Theory

- How good are classifiers?
  - Flexible
  - Better than BT models
  - Robust to annotation noises and data scarcity (more in paper)



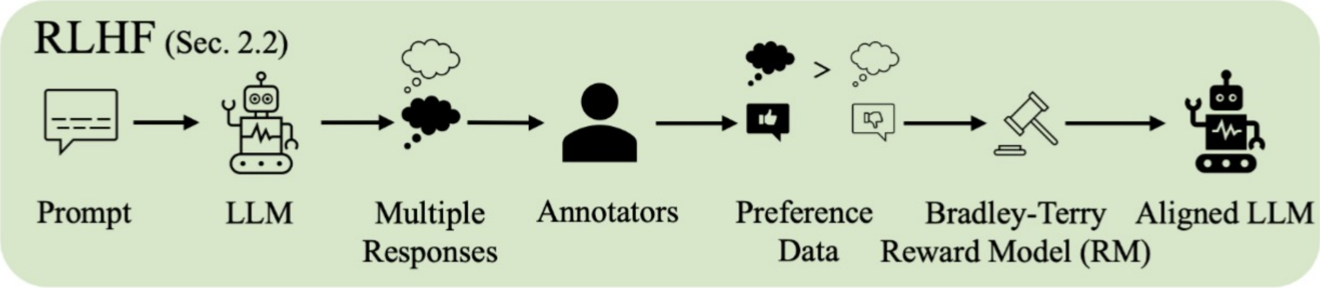
van_der_Schaar \ LAB

vanderschaar-lab.com

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

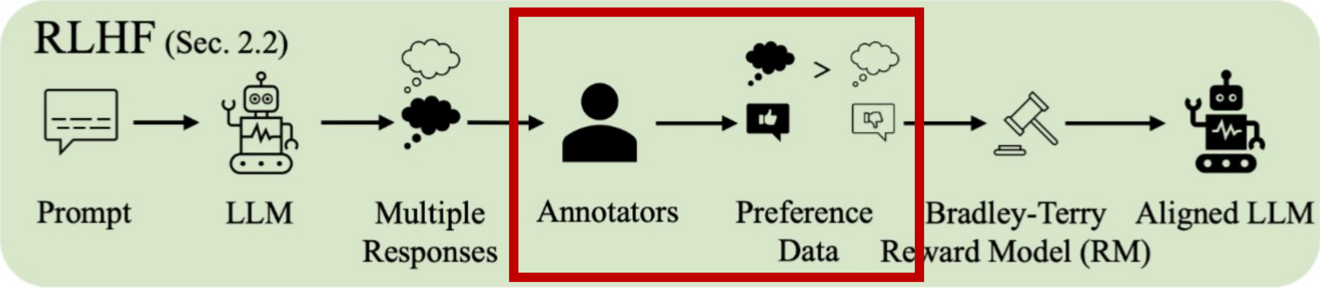
- Common practice of alignment: RLHF



- Why do we need to align LLMs from demonstrations?

RM from Demonstration: Inverse RL for Alignment

- Common practice of alignment: RLHF

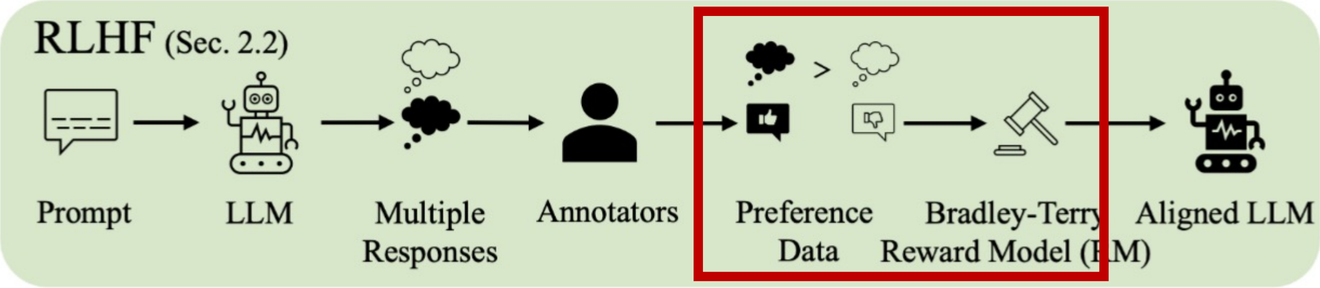


- Why do we need to align LLMs from demonstrations?
  - 1. Preference-based alignment is expensive

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Common practice of alignment: RLHF

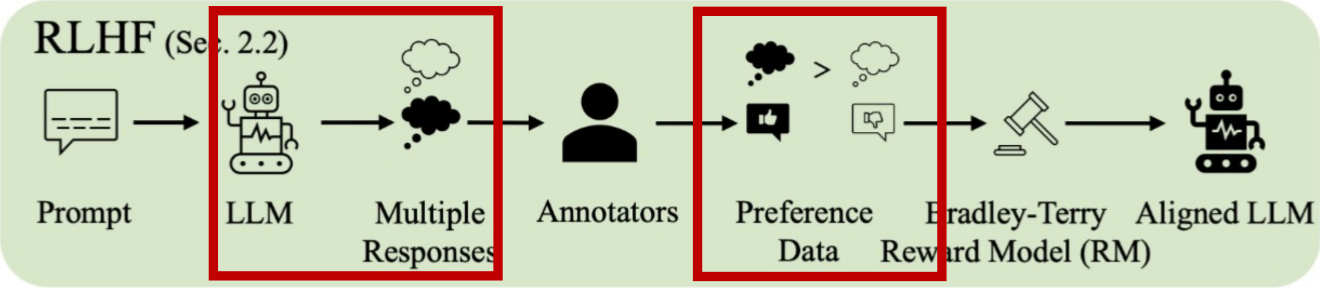


- Why do we need to align LLMs from demonstrations?
  - 1. Preference-based alignment is expensive
  - 2. Assumptions such as Bradley-Terry models are needed

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Common practice of alignment: RLHF

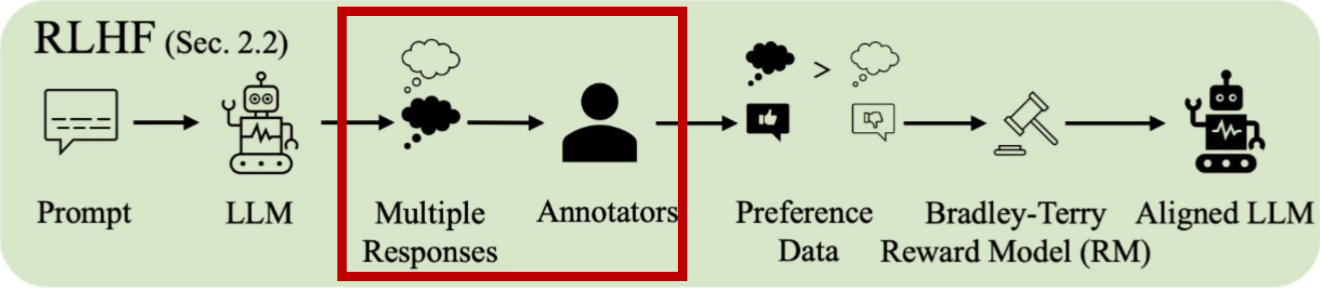


- Why do we need to align LLMs from demonstrations?
  - 1. Preference-based alignment is expensive
  - 2. Assumptions such as Bradley-Terry models are needed
  - 3. Noisy preference annotations

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Common practice of alignment: RLHF

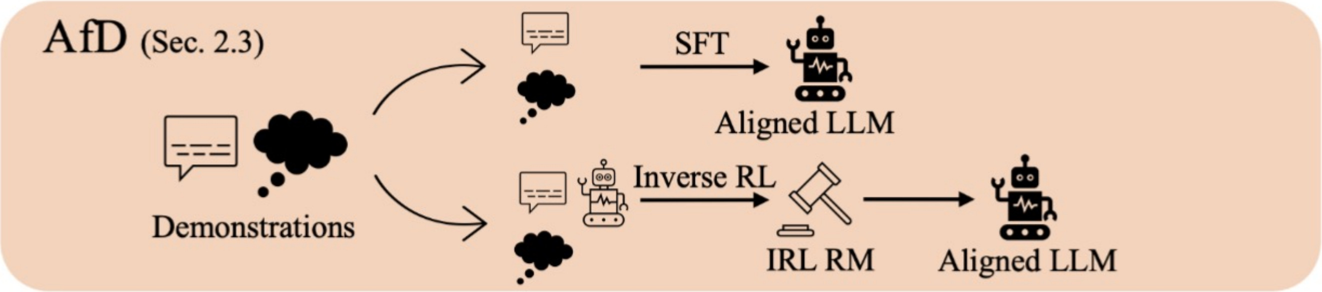


- Why do we need to align LLMs from demonstrations?
  - 1. Preference-based alignment is expensive
  - 2. Assumptions such as Bradley-Terry models are needed
  - 3. Noisy preference annotations
  - 4. Privacy concerns

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

RM from Demonstration: Inverse RL for Alignment

- Our solution: Alignment from Demonstrations using Inverse RL

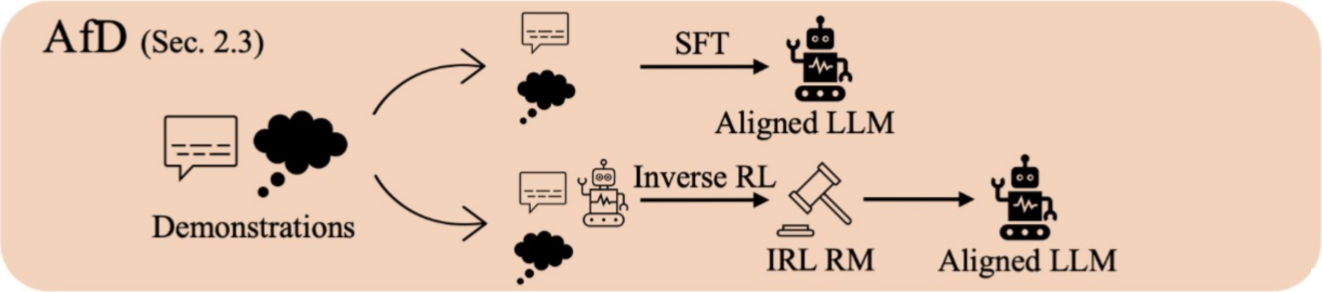


e.g., Prescribe a medicine
Stylize…

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

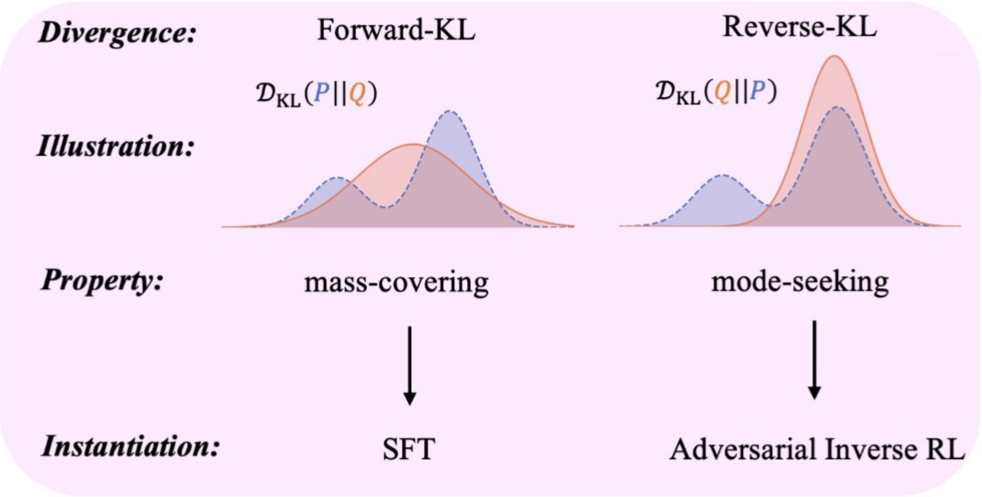
RM from Demonstration: Inverse RL for Alignment

- Our solution: Alignment from Demonstrations using Inverse RL



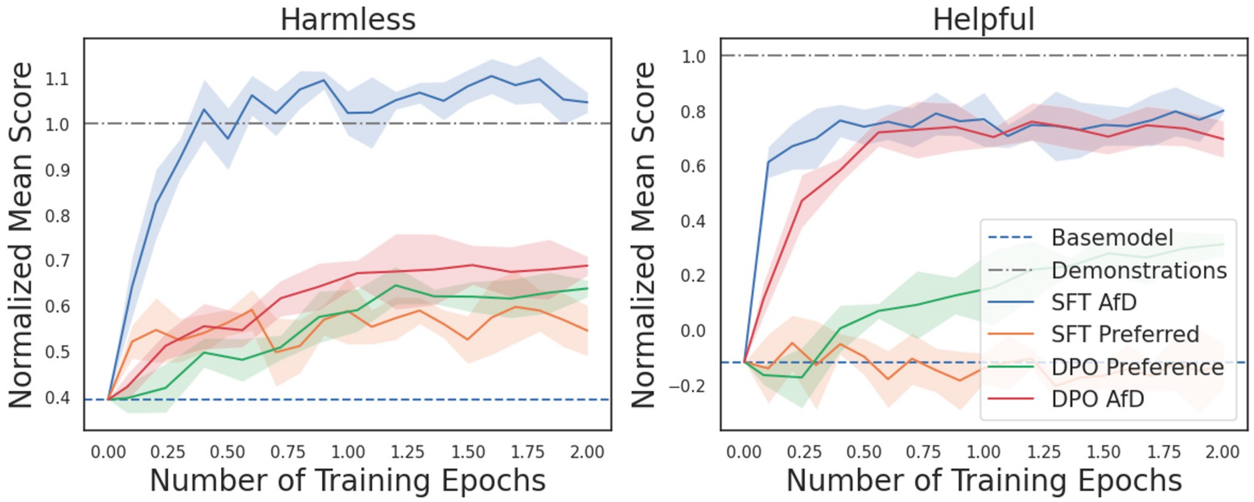
e.g., Prescribe a medicine
Stylize…

- From an IRL perspective
  - Distributional matching
  - SFT = Forward KL for distribution matching
  - Reverse KL? requires (smart) reward modeling



RM from Demonstration: Inverse RL for Alignment

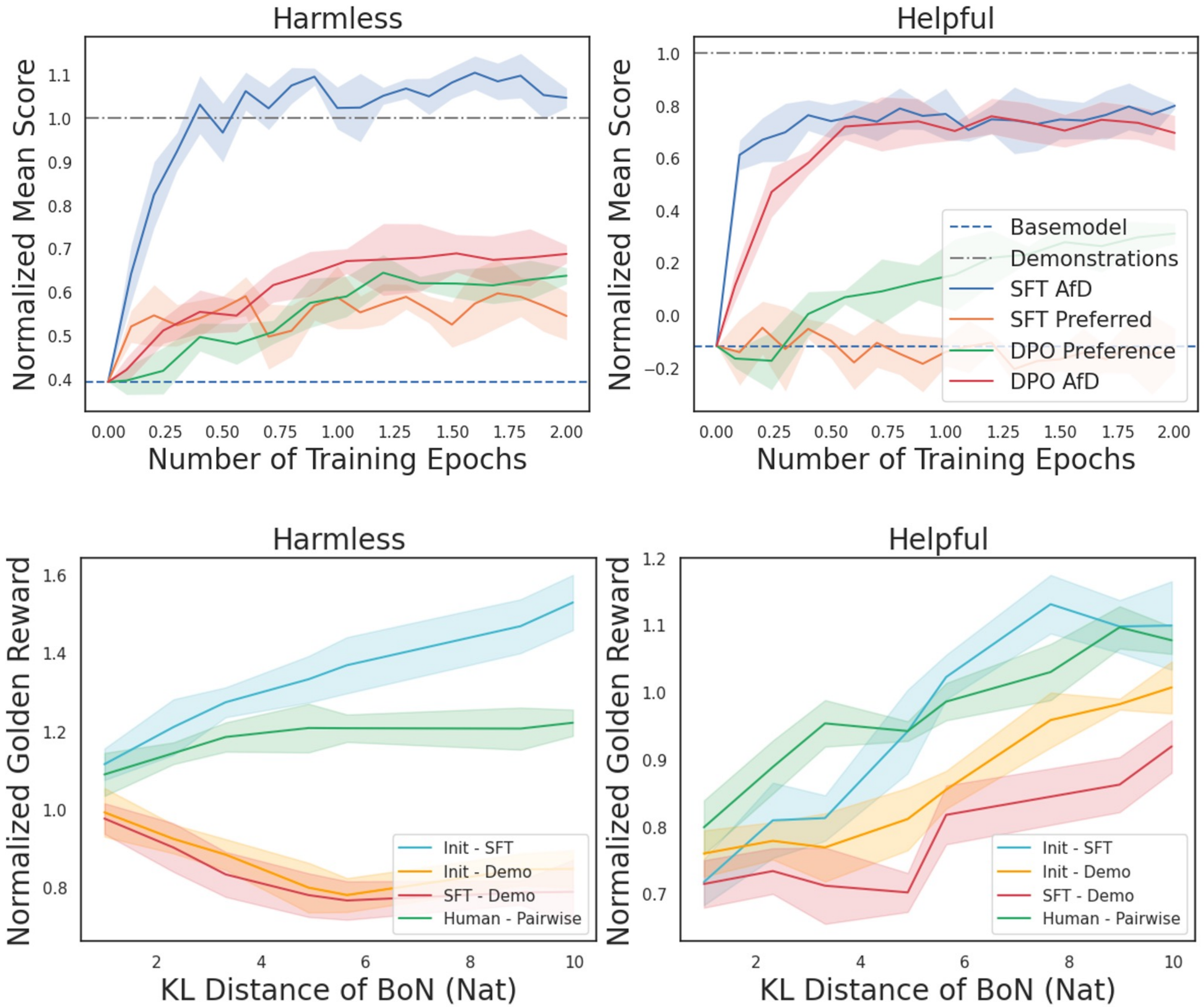
- Forward-KL can be strong enough



RM from Demonstration: Inverse RL for Alignment

- Forward-KL can be strong enough

- Reverse-KL (reward modeling) further improves performance



van_der_Schaar
\ LAB

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE

Thank you!

References

[Position] *Improving LLM Generation with Inverse and Forward Alignment: Reward Modeling, Prompting, Fine-Tuning, and Inference-Time Optimization*
NeurIPS'2024 System2 Reasoning workshop
Hao Sun, Thomas Pouplin, Nicolás Astorga, Tennison Liu, Mihaela van der Schaar

[Prompt-OIRL] *Query-Dependent Prompt Evaluation and Optimization with Offline Inverse RL*
ICLR'2024
Hao Sun, Alihan Hüyük, Mihaela van der Schaar

[DenseReward] *Dense Reward for Free in Reinforcement Learning from Human Feedback*
ICML'2024
Alex Chan, **Hao Sun**, Samuel Holt, Mihaela van der Schaar

[DataCOPE] *When is Off-Policy Evaluation (Reward Modeling) Useful in Contextual Bandits? A Data-Centric Perspective*
Journal of Data-Centric Machine Learning Research (DMLR)
Hao Sun*, Alex Chan*, Nabeel Seedat, Alihan Hüyük, Mihaela van der Schaar

[InverseRLignment] *Inverse-RLignment: Inverse Reinforcement Learning from Demonstrations for LLM Alignment*
RLC'2024 RL Beyond Reward workshop
Hao Sun and Mihaela van der Schaar

[RATP] *Retrieval Augmented Thought Process for Private Data Handling in Healthcare*
Preprint
Thomas Pouplin*, **Hao Sun***, Samuel Holt, Mihaela van der Schaar

[RMBeyondBT] Rethinking the Bradley-Terry Models in Preference-based Reward Modeling: Foundation, Theory, and its Alternatives
Preprint
Hao Sun*, Yunyi Shen*, Jean-Francois Ton

hs789@cam.ac.uk
UNIVERSITY OF CAMBRIDGE