Dynamic Rewarding with Prompt Optimization Enables Tuning-free Self-Alignment of Language Models

Somanshu Singla[1*], Zhen Wang[12*], Tianyang Liu[1], Abdullah Ashfaq[1], Zhiting Hu[1], Eric P. Xing[23]

[1]UC San Diego, [2]MBZUAI, [3]CMU

EMNLP 2024

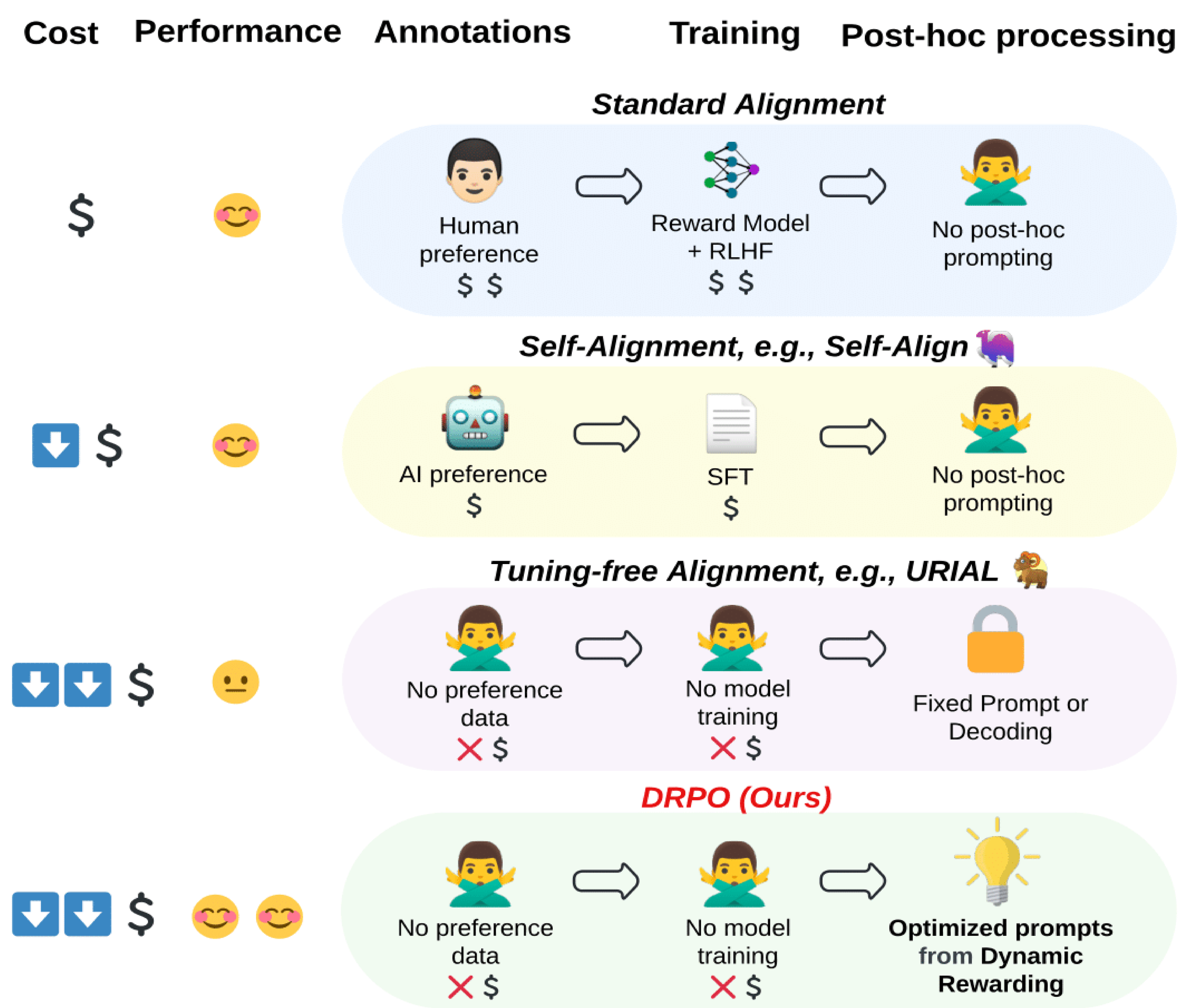
LLM Alignment is expensive

Traditional alignment methods (RLHF/SFT) are:

- Effective and achieve great performance in SoTA LLMs.
- **But** resource-intensive and need extensive human annotations

We need more **cost-efficient and high-performance methods:**

- **Self-alignment:** aligning LLMs by themselves, less annotations
- **Tuning-free alignment**: inference-time alignment, no training cost



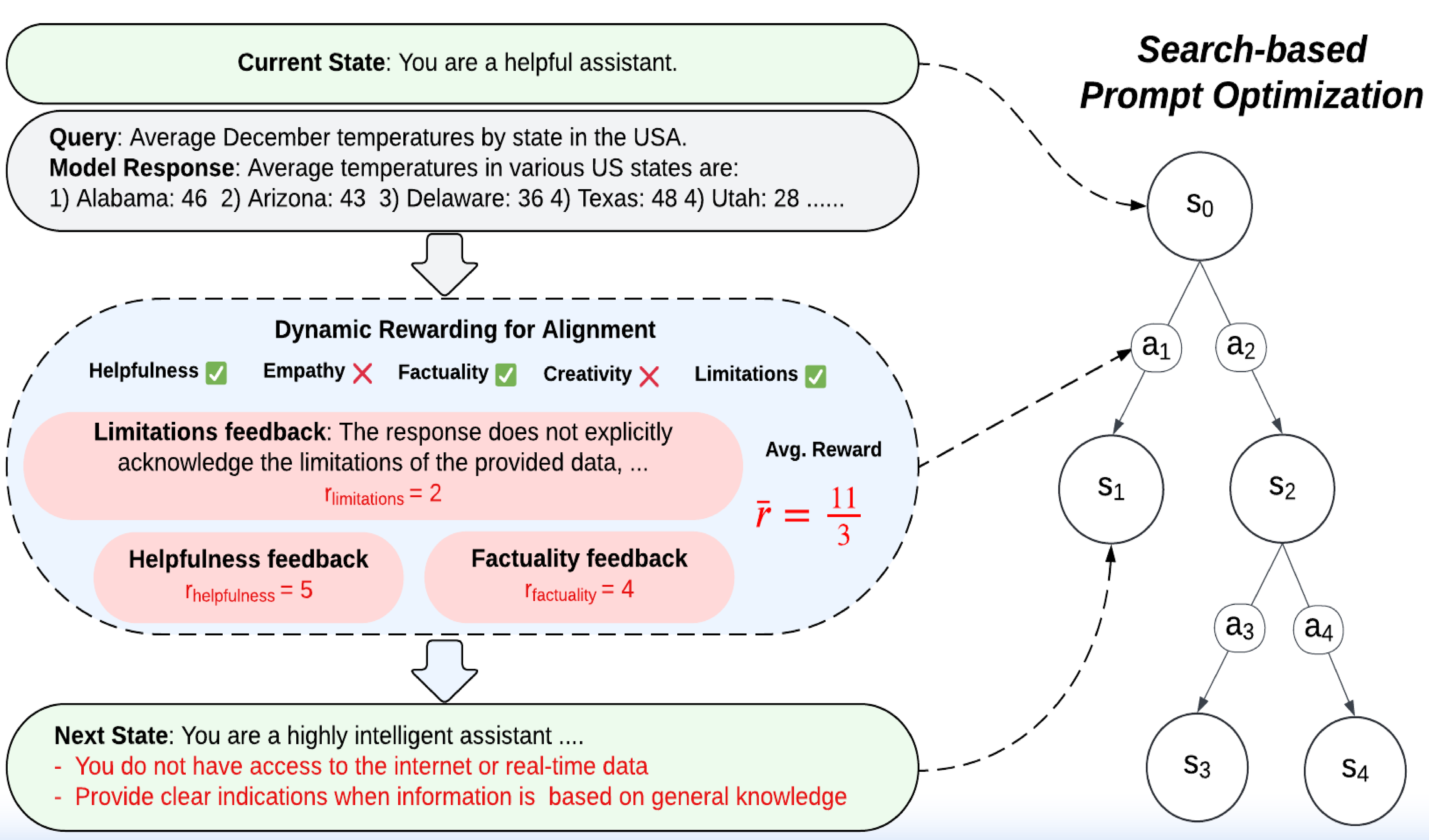
However, **self-alignment** still requires tuning and some annotations; **Tuning-free align.** tends to be static, relying on fixed rewards/prompts

DRPO: Tuning-free Self-alignment

Goal: Design a tuning-free self-alignment method without relying on humans, with great generalizability across various LLMs.

Key innovations (check more details in the paper):

1. Inference-time optimization with Dynamic Rewarding (DR)
2. DR provides dynamic feedback for model-specific alignment
3. DRPO generalizes across LLMs with no training and annotations



Superior Alignment Unlocked

1. Superior alignment performance when compared with various RLHF/tuning-free methods on *just-eval-instruct* benchmark

[Tuned] Model	Method	K	Helpful	Clear	Factual	Deep	Engage	Avg.
[✗] Mistral 7b	Base	0	2.20	2.51	2.29	1.69	1.80	2.10
[✗] Mistral 7b	URIAL	3	3.62	4.32	3.75	2.70	3.41	3.56
[✗] Mistral 7b	DRPO	2	**4.23**	**4.56**	**3.97**	**3.68**	**3.84**	**4.06**
[✓] Mistral 7b (Instruct)	Base	0	3.98	4.44	3.64	2.97	3.26	3.66
[✓] Mistral 7b (Instruct)	URIAL	3	3.94	4.51	3.69	2.99	3.75	3.78
[✓] Mistral 7b (Instruct)	DRPO	2	**4.22**	**4.60**	**3.80**	**3.68**	**3.99**	**4.06**
[✗] Llama 2 70b^q	Base	0	2.07	2.55	2.35	1.50	1.63	2.02
[✗] Llama 2 70b^q	URIAL	3	4.25	4.67	4.03	3.08	3.80	3.97
[✗] Llama 2 70b^q	DRPO	2	**4.42**	**4.72**	**4.23**	**3.81**	**3.98**	**4.23**
[✓] Llama 2 70b^q (chat)	Base	0	4.36	4.71	3.95	3.56	3.76	4.07
[✓] Llama 2 70b^q (chat)	URIAL	3	4.32	4.72	4.08	3.50	4.25	4.17
[✓] Llama 2 70b^q (chat)	DRPO	2	**4.46**	**4.75**	**4.10**	**4.11**	**4.37**	**4.36**
[✗] Llama 3 8b	Base	0	1.82	2.27	2.20	1.38	1.48	1.83
[✗] Llama 3 8b	URIAL	3	3.94	**4.51**	3.69	2.99	**3.75**	3.78
[✗] Llama 3 8b	DRPO	2	**4.02**	4.40	**3.84**	**3.50**	3.65	**3.88**
[✓] Llama 3 8b (Instruct)	Base	0	4.43	4.72	3.98	3.45	3.76	4.07
[✓] Llama 3 8b (Instruct)	URIAL	3	4.48	4.81	**4.19**	3.55	4.27	4.26
[✓] Llama 3 8b (Instruct)	DRPO	2	**4.54**	**4.81**	4.16	**4.08**	**4.40**	**4.40**
[✓] gpt-3.5-turbo	Base	0	4.56	4.89	4.41	3.30	3.55	4.14
[✓] gpt-3.5-turbo	URIAL	3	4.30	4.77	4.41	3.44	4.11	4.21
[✓] gpt-3.5-turbo	DRPO	2	**4.67**	**4.92**	**4.53**	**4.07**	**4.58**	**4.55**
[✓] gpt-4-0613	Base	0	**4.71**	**4.93**	**4.52**	3.49	3.53	**4.24**

2. Alignment prompt transfer across various base models

Model	Mistral Prompt	Llama Prompt	Base Prompt
Mistral 7b	**4.06**	4.03	4.04
Llama 2 70b^q	4.19	**4.23**	4.17

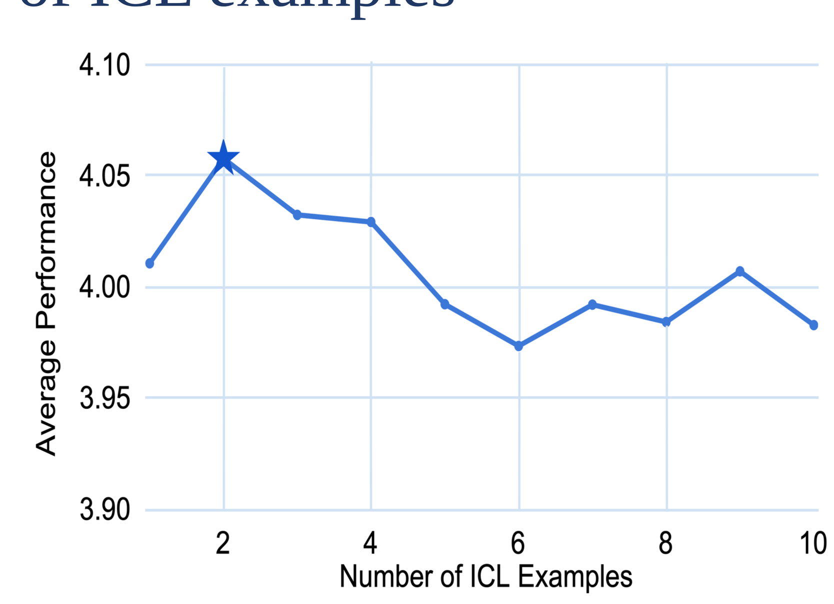
3. Ablation study showing the power of dynamic rewarding

Model	Dynamic Reward Prompt	Dynamic Reward ICL	Avg.
Mistral 7b (Instruct)	✓	✓	**4.06**
Mistral 7b (Instruct)	✗	✓	4.02
Mistral 7b (Instruct)	✓	✗	3.86

4. Ablation study showing the importance of both alignment prompt and ICL examples

Model	System Prompt	ICL ($K = 2$)	Avg.
Mistral 7b	✓	✓	**4.06**
Mistral 7b (Instruct)	✓	✓	**4.06**
Llama 2 70b^q	✓	✓	**4.23**
gpt-3.5-turbo	✓	✓	**4.55**
Mistral 7b	✗	✓	4.04
Mistral 7b (Instruct)	✗	✓	4.04
Llama 2 70b^q	✗	✓	4.17
gpt-3.5-turbo	✗	✓	4.42
Mistral 7b (Instruct)	✓	✗	3.67
Llama 2 70b^q	✓	✗	3.63
gpt-3.5-turbo	✓	✗	4.34

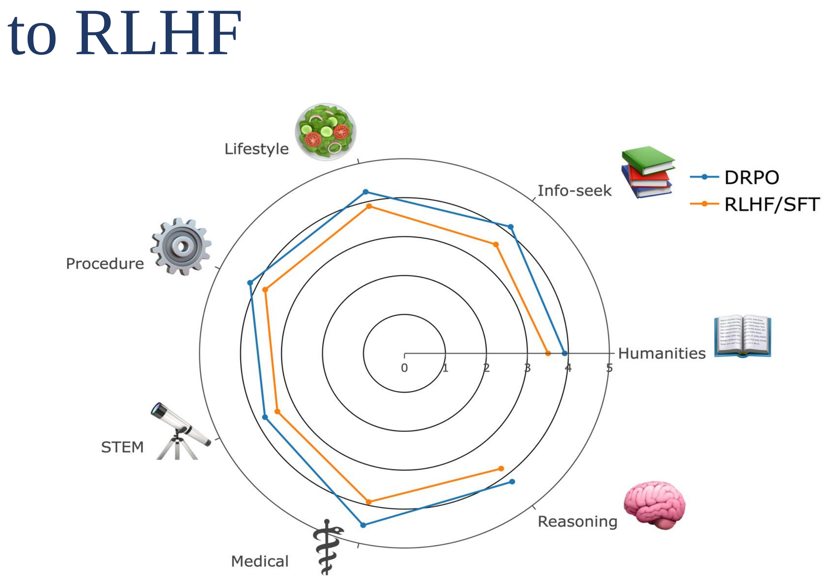
5. Alignment performance of Mistral-7B with varying number of ICL examples



6. Optimized alignment prompts, showing model-specific insights (highlighted colors) for **gpt-3.5-turbo**

> As a helpful and ethical assistant, your primary goal is to provide responses that are accurate, engaging, clear, and emotionally resonant across a wide range of queries.
> - Strive to make complex topics understandable and emotionally engaging, communicating in a human-like and relatable manner. Organize your responses to enhance readability and emotional connection, avoiding overly technical jargon.
> - Always acknowledge the limitations of your knowledge, especially when speculating about historical 'what-ifs', future predictions, or interpreting emotions.
> - Aim for a balance between detailed, informative content and a conversational, engaging tone. Incorporate storytelling elements, examples, analogies, and direct questions to make information relatable.
> - Avoid overwhelming the user with excessive information; structure your responses to be clear, well-organized, and mindful of the user's cognitive load.

7. Categorized performance Mistral-7B of DRPO compared to RLHF



Insights: DRPO identifies model-specific alignment weaknesses:

1. It recognizes that model tends to overcomplicate things.
2. Provides actionable insights such storytelling etc.

Road Ahead with DRPO

- DRPO is capable of finding **model specific weaknesses** and can be utilized towards data efficient alignment.
- **Dynamic rewarding** will be applied and studied on a wide variety of problems such as **Judge Models, Agentic Systems**, etc.