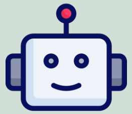
Understanding and Overcoming Pitfalls in Language Model Alignment

Noam Razin

Princeton Language and Intelligence, Princeton University

Language Models

Language Model (LM): Neural network trained to produce a **distribution over text**

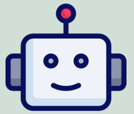


LM π_θ

θ - Parameters of the LM

Language Models

Language Model (LM): Neural network trained to produce a **distribution over text**



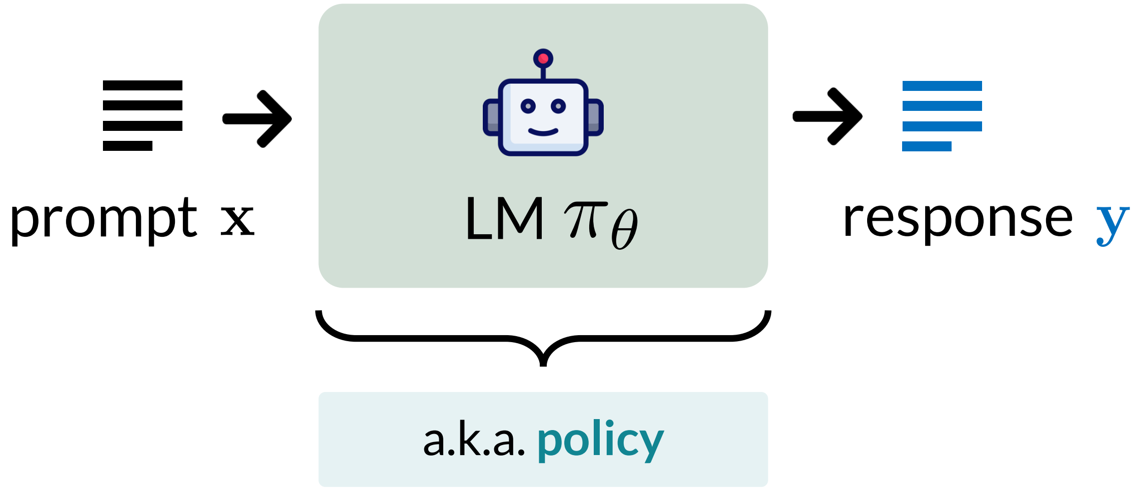
LM π_θ

a.k.a. **policy**

θ - Parameters of the LM

Language Models

Language Model (LM): Neural network trained to produce a **distribution over text**



θ - Parameters of the LM

Supervised Finetuning of LMs

To ensure LMs generate safe and helpful content, they are aligned via **finetuning**

Wei et al. 2022, Chung et al. 2022, Ouyang et al. 2022, Zhang et al. 2023

Supervised Finetuning of LMs

To ensure LMs generate safe and helpful content, they are aligned via **finetuning**

Supervised Finetuning (SFT)

Minimize a standard next-token prediction loss over **desired responses**

prompt $\mathbf{x}$ desired response $\mathbf{y}$

Wei et al. 2022, Chung et al. 2022, Ouyang et al. 2022, Zhang et al. 2023

Supervised Finetuning of LMs

To ensure LMs generate safe and helpful content, they are aligned via **finetuning**

Supervised Finetuning (SFT)

Minimize a standard next-token prediction loss over **desired responses**

prompt $\mathbf{x}$ desired response $\mathbf{y}$

Limitation of SFT:

Hard to formalize human preferences through labels

Wei et al. 2022, Chung et al. 2022, Ouyang et al. 2022, Zhang et al. 2023

Finetuning LMs via Preference Data

Finetuning LMs via Preference Data

Preference-Based Finetuning

Limitations of SFT led to wide adoption of approaches using **preference data**

prompt $\mathbf{x}$

preferred response $\mathbf{y}^+$

dispreferred response $\mathbf{y}^-$

Finetuning LMs via Preference Data

Preference-Based Finetuning

Limitations of SFT led to wide adoption of approaches using **preference data**

prompt $\mathbf{x}$ | preferred response $\mathbf{y}^+$ | dispreferred response $\mathbf{y}^-$

Underlying Assumption: Preferences are governed by an **unknown ground truth reward**

$$r_{\mathrm{G}}(\mathbf{x}, \mathbf{y}^+) > r_{\mathrm{G}}(\mathbf{x}, \mathbf{y}^-)$$

Finetuning LMs via Preference Data

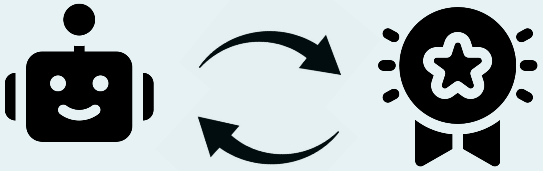
Q: How can we maximize r_{G} if we only have access to it through preference data?

Finetuning LMs via Preference Data

Q: How can we maximize r_{G} if we only have access to it through preference data?

Reinforcement Learning

(e.g. Ouyang et al. 2022)

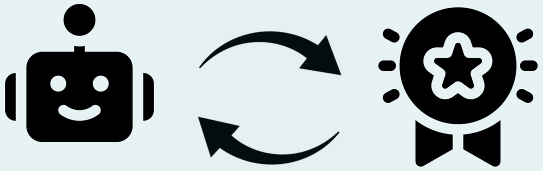


Finetuning LMs via Preference Data

Q: How can we maximize r_G if we only have access to it through preference data?

Reinforcement Learning

(e.g. Ouyang et al. 2022)



Direct Preference Learning

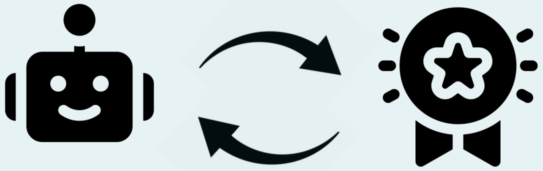
(e.g. Rafailov et al. 2023)

Finetuning LMs via Preference Data

Q: How can we maximize r_{G} if we only have access to it through preference data?

Reinforcement Learning

(e.g. Ouyang et al. 2022)



Direct Preference Learning

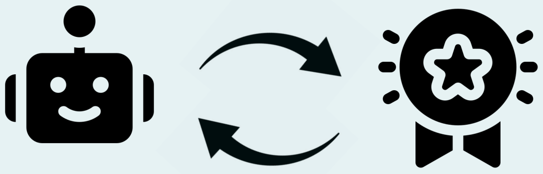
(e.g. Rafailov et al. 2023)

We Will See: Limited understanding can lead to undesirable outcomes

Part I: Alignment via Reinforcement Learning

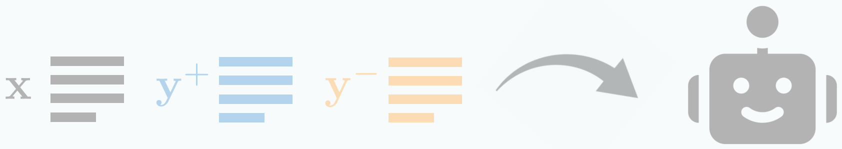
Reinforcement Learning

(e.g. Ouyang et al. 2022)



Direct Preference Learning

(e.g. Rafailov et al. 2023)



Vanishing Gradients in Reinforcement Finetuning of Language Models

R + Zhou + Saremi + Thilak + Bradley + Nakkiran + Susskind + Littwin | *ICLR 2024*

What Makes a Reward Model a Good Teacher? An Optimization Perspective

R + Wang + Strauss + Wei + Lee + Arora | *arXiv 2025*

Why is Your Language Model a Poor Implicit Reward Model?

R + Lin + Yao + Arora | *arXiv 2025*

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF)

Schulman et al. 2017, Christiano et al. 2017, Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF)

1. Learn a proxy **reward model (RM)** $r_{\mathrm{RM}}(\mathbf{x}, \mathbf{y})$ by fitting preference data

$$\mathbf{x} \qquad \mathbf{y}^{+} \qquad \mathbf{y}^{-}$$

Schulman et al. 2017, Christiano et al. 2017, Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF)

1. Learn a proxy **reward model (RM)** $r_{\text{RM}}(\mathbf{x}, \mathbf{y})$ by fitting preference data

2. Maximize proxy reward via **policy gradient methods** (e.g. PPO) over set of prompts $\mathcal{S}$

Schulman et al. 2017, Christiano et al. 2017, Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF)

1. Learn a proxy **reward model (RM)** $r_{\mathrm{RM}}(\mathbf{x}, \mathbf{y})$ by fitting preference data

$\mathbf{x}$ $\quad$ $\mathbf{y}^+$ $\quad$ $\mathbf{y}^-$

2. Maximize proxy reward via **policy gradient methods** (e.g. PPO) over set of prompts $\mathcal{S}$

$$\phi_{\mathrm{RLHF}}(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\Big[\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x}, \mathbf{y})\right] - \lambda \cdot \mathrm{KL}\big(\pi_\theta(\cdot|\mathbf{x})||\pi_{\mathrm{ref}}(\cdot|\mathbf{x})\big)\Big]$$

Schulman et al. 2017, Christiano et al. 2017, Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF)

1. Learn a proxy **reward model (RM)** $r_{\mathrm{RM}}(\mathbf{x}, \mathbf{y})$ by fitting preference data

$$\mathbf{x} \qquad \mathbf{y}^+ \qquad \mathbf{y}^-$$

2. Maximize proxy reward via **policy gradient methods** (e.g. PPO) over set of prompts $\mathcal{S}$

$$\phi_{\mathrm{RLHF}}(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\left[\underbrace{\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x}, \mathbf{y})\right]}_{\text{maximize reward}} - \lambda \cdot \mathrm{KL}\big(\pi_\theta(\cdot|\mathbf{x}) || \pi_{\mathrm{ref}}(\cdot|\mathbf{x})\big)\right]$$

Schulman et al. 2017, Christiano et al. 2017, Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF)

1. Learn a proxy **reward model (RM)** $r_{\mathrm{RM}}(\mathbf{x}, \mathbf{y})$ by fitting preference data

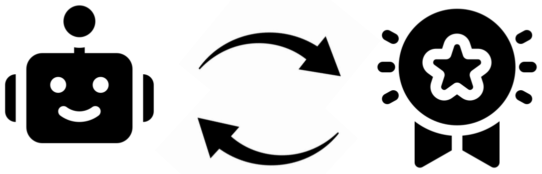
2. Maximize proxy reward via **policy gradient methods** (e.g. PPO) over set of prompts $\mathcal{S}$

$$\phi_{\mathrm{RLHF}}(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\Big[\underbrace{\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})\right]}_{\text{maximize reward}} - \lambda\cdot\underbrace{\mathrm{KL}\big(\pi_\theta(\cdot|\mathbf{x})||\pi_{\mathrm{ref}}(\cdot|\mathbf{x})\big)}_{\text{stay close to initial policy}}\Big]$$

Schulman et al. 2017, Christiano et al. 2017, Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022

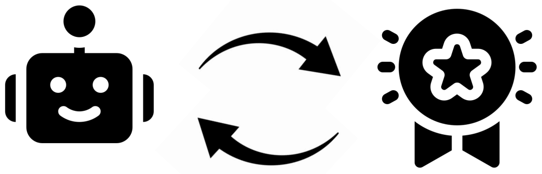
Evaluating RMs

The success of RLHF depends heavily on the **quality of the RM**



Evaluating RMs

The success of RLHF depends heavily on the **quality of the RM**



But it is unclear how we should **evaluate this quality**

Evaluating RMs

Currently, RMs are primarily evaluated through **accuracy**

Evaluating RMs

Currently, RMs are primarily evaluated through **accuracy**

Definition: Accuracy

For prompt $\mathbf{x}$ and distribution $\mathcal{D}$ over pairs $\{\mathbf{y}, \mathbf{y}'\}$:

$$\mathbb{E}_{\{\mathbf{y},\mathbf{y}'\}\sim\mathcal{D}}\left[\mathbb{1}\left[r_{\mathrm{RM}} \text{ ranks } \mathbf{y}, \mathbf{y}' \text{ the same as } r_{\mathrm{G}}\right]\right]$$

Evaluating RMs

Currently, RMs are primarily evaluated through **accuracy**

Definition: Accuracy

For prompt $\mathbf{x}$ and distribution $\mathcal{D}$ over pairs $\{\mathbf{y}, \mathbf{y}'\}$:

$$\mathbb{E}_{\{\mathbf{y},\mathbf{y}'\}\sim\mathcal{D}}\left[\mathbb{1}\left[r_{\mathrm{RM}} \text{ ranks } \mathbf{y}, \mathbf{y}' \text{ the same as } r_{\mathrm{G}}\right]\right]$$

RewardBench Lambert et al. 2024

	Model	Score
1	infly/INF-ORM-Llama3.1-70B	95.1
2	ShikaiChen/LDL-Reward-Gemma-2-27B-v0.1	95.0
3	nicolinho/QRM-Gemma-2-27B	94.4

Evaluating RMs

Currently, RMs are primarily evaluated through **accuracy**

Definition: Accuracy

For prompt $\mathbf{x}$ and distribution $\mathcal{D}$ over pairs $\{\mathbf{y}, \mathbf{y}'\}$:

$$\mathbb{E}_{\{\mathbf{y},\mathbf{y}'\}\sim\mathcal{D}}\left[\mathbb{1}\left[r_{\mathrm{RM}} \text{ ranks } \mathbf{y}, \mathbf{y}' \text{ the same as } r_{\mathrm{G}}\right]\right]$$

rewardbench Lambert et al. 2024

	Model	Score
1	infly/INF-ORM-Llama3.1-70B	95.1
2	ShikaiChen/LDL-Reward-Gemma-2-27B-v0.1	95.0
3	nicolinho/QRM-Gemma-2-27B	94.4

Intuitively, accuracy quantifies the extent to which maximizing r_{RM} is likely to increase r_{G}

Are More Accurate RMs Always Better?

Q: Are more accurate reward models better teachers for RLHF?

Are More Accurate RMs Always Better?

Q: Are more accurate reward models better teachers for RLHF?

❌

Not necessarily!

Main Contributions: What Makes a Good RM

Main Contributions: What Makes a Good RM

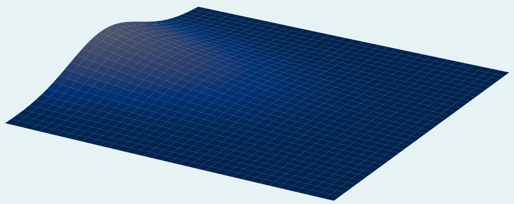
Optimization Perspective: When does an RM enable efficient policy gradient optimization?

Main Contributions: What Makes a Good RM

Optimization Perspective: When does an RM enable efficient policy gradient optimization?

1

Regardless of how accurate the RM is, it can **induce a flat objective landscape** that hinders optimization

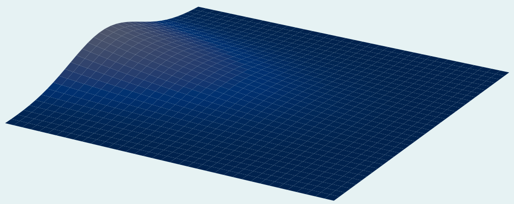


Main Contributions: What Makes a Good RM

Optimization Perspective: When does an RM enable efficient policy gradient optimization?

1

Regardless of how accurate the RM is, it can **induce a flat objective landscape** that hinders optimization



2

Implication I:

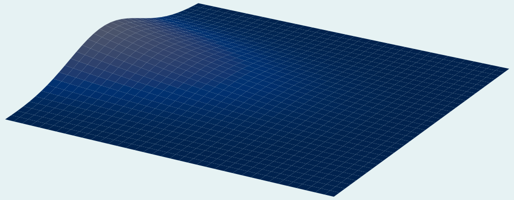
More accurate RMs are not necessarily better teachers for RLHF

Main Contributions: What Makes a Good RM

Optimization Perspective: When does an RM enable efficient policy gradient optimization?

1

Regardless of how accurate the RM is, it can **induce a flat objective landscape** that hinders optimization

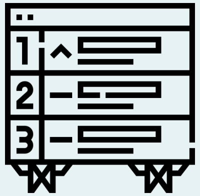


2

Implication I:
More accurate RMs are not necessarily better teachers for RLHF

3

Implication II:
Fundamental limitations of existing RM benchmarks



Reward Variance

Definition: Reward Variance

The reward variance that r_{RM} induces for π_θ and $\mathbf{x}$ is:

$$\mathrm{Var}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})] :=$$

Reward Variance

Definition: Reward Variance

The reward variance that r_{RM} induces for π_θ and $\mathbf{x}$ is:

$$\mathrm{Var}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})] := \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\left(r_{\mathrm{RM}}(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathbf{y}'\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y}')\right]\right)^2\right]$$

Reward Variance

Definition: Reward Variance

The reward variance that r_{RM} induces for π_θ and $\mathbf{x}$ is:

$$\mathrm{Var}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})] := \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\left(r_{\mathrm{RM}}(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathbf{y}'\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y}')\right]\right)^2\right]$$

Interpretation: Reward variance measures how well r_{RM} separates responses that are probable under π_θ

Reward Variance

Definition: Reward Variance

The reward variance that r_{RM} induces for π_θ and $\mathbf{x}$ is:

$$\text{Var}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}[r_{\text{RM}}(\mathbf{x},\mathbf{y})] := \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\left(r_{\text{RM}}(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathbf{y}'\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\text{RM}}(\mathbf{x},\mathbf{y}')\right]\right)^2\right]$$

Interpretation: Reward variance measures how well r_{RM} separates responses that are probable under π_θ

In Contrast: Accuracy depends only on how r_{RM} ranks different responses

Low Reward Variance Implies Slow Reward Maximization

Low Reward Variance Implies Slow Reward Maximization

Theorem

The time it takes for the expected reward, measured w.r.t. any reward function, to increase by an additive constant is:

Low Reward Variance Implies Slow Reward Maximization

Theorem

The time it takes for the expected reward, measured w.r.t. any reward function, to increase by an additive constant is:

$$\Omega\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\left[\mathrm{Var}_{\mathbf{y}\sim\pi_{\mathrm{ref}}(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})\right]\right]^{-\frac{1}{3}}\right)$$

Low Reward Variance Implies Slow Reward Maximization

Theorem

The time it takes for the expected reward, measured w.r.t. any reward function, to increase by an additive constant is:

$$\Omega\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\left[\mathrm{Var}_{\mathbf{y}\sim\pi_{\mathrm{ref}}(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})\right]\right]^{-\frac{1}{3}}\right)$$

Proof Idea: The gradient and Hessian vanish when reward variance is low

Low Reward Variance Implies Slow Reward Maximization

Theorem

The time it takes for the expected reward, measured w.r.t. any reward function, to increase by an additive constant is:

$$\Omega\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\left[\mathrm{Var}_{\mathbf{y}\sim\pi_{\mathrm{ref}}(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})\right]\right]^{-\frac{1}{3}}\right)$$

Proof Idea: $\underbrace{\text{The gradient and Hessian vanish when reward variance is low}}_{\text{holds for any RL setting with softmax policies (not just LMs)}}$

Low Reward Variance Implies Slow Reward Maximization

Theorem

The time it takes for the expected reward, measured w.r.t. any reward function, to increase by an additive constant is:

$$\Omega\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\left[\mathrm{Var}_{\mathbf{y}\sim\pi_{\mathrm{ref}}(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})\right]\right]^{-\frac{1}{3}}\right)$$

Proof Idea: $\underbrace{\text{The gradient and Hessian vanish when reward variance is low}}_{\text{holds for any RL setting with softmax policies (not just LMs)}}$

RM needs to induce sufficient variance for efficient optimization

Implication I: More Accurate RMs Are Not Necessarily Better

Implication I: More Accurate RMs Are Not Necessarily Better

Theorem

For any initial policy π_{ref}, there exist a **perfectly accurate** r_{per} and **relatively inaccurate** r_{inacc} such that:

Implication I: More Accurate RMs Are Not Necessarily Better

Theorem

For any initial policy π_{ref}, there exist a **perfectly accurate** r_{per} and **relatively inaccurate** r_{inacc} such that:

$\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\text{G}}(\mathbf{x},\mathbf{y})\right]$ increases **arbitrarily slower** when training with r_{per} compared to r_{inacc}

Implication I: More Accurate RMs Are Not Necessarily Better

Theorem

For any initial policy π_{ref}, there exist a **perfectly accurate** r_{per} and **relatively inaccurate** r_{inacc} such that:

$\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\text{G}}(\mathbf{x},\mathbf{y})\right]$ increases **arbitrarily slower** when training with r_{per} compared to r_{inacc}

*Same holds with almost any accuracy values for the RMs

Illustration: Effect of Accuracy and Reward Variance

Illustration: Effect of Accuracy and Reward Variance

Ground Truth Reward

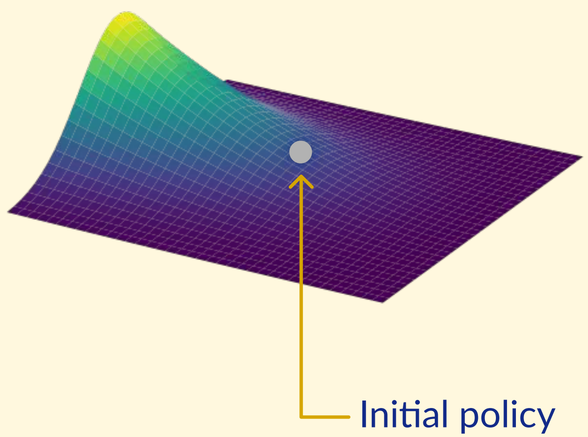
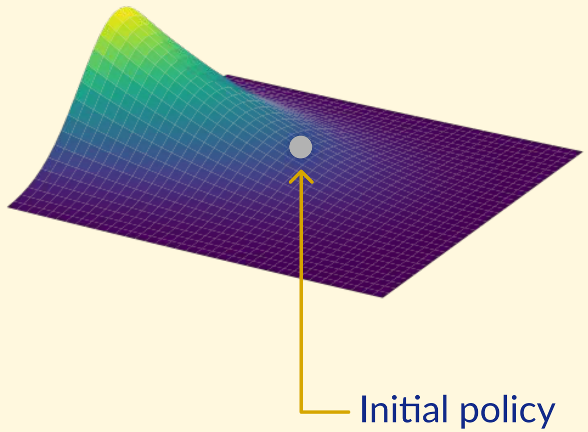


Illustration: Effect of Accuracy and Reward Variance

Ground Truth Reward

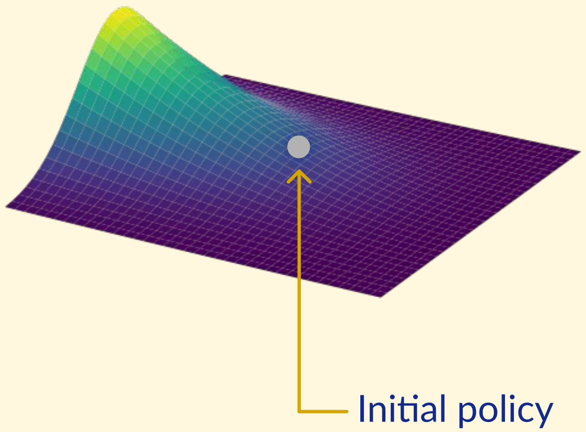


RM

Accuracy and reward variance capture distinct aspects of an RM

Illustration: Effect of Accuracy and Reward Variance

Ground Truth Reward



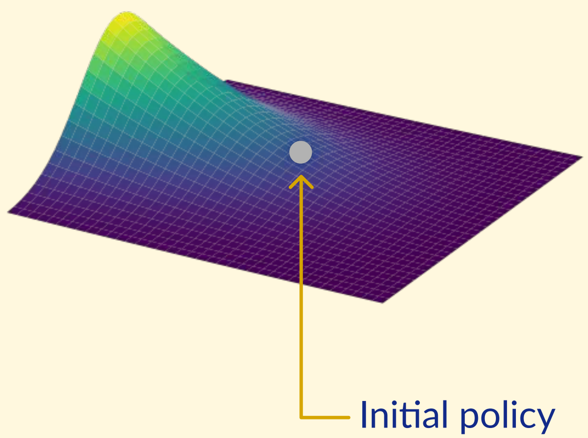
RM

Accuracy and reward variance capture distinct aspects of an RM

Accuracy

Illustration: Effect of Accuracy and Reward Variance

Ground Truth Reward

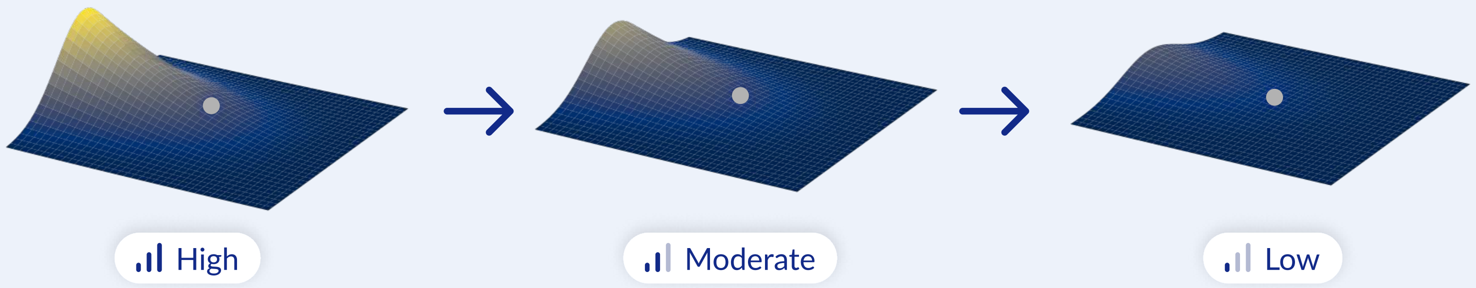


RM

Accuracy and reward variance capture distinct aspects of an RM

Accuracy

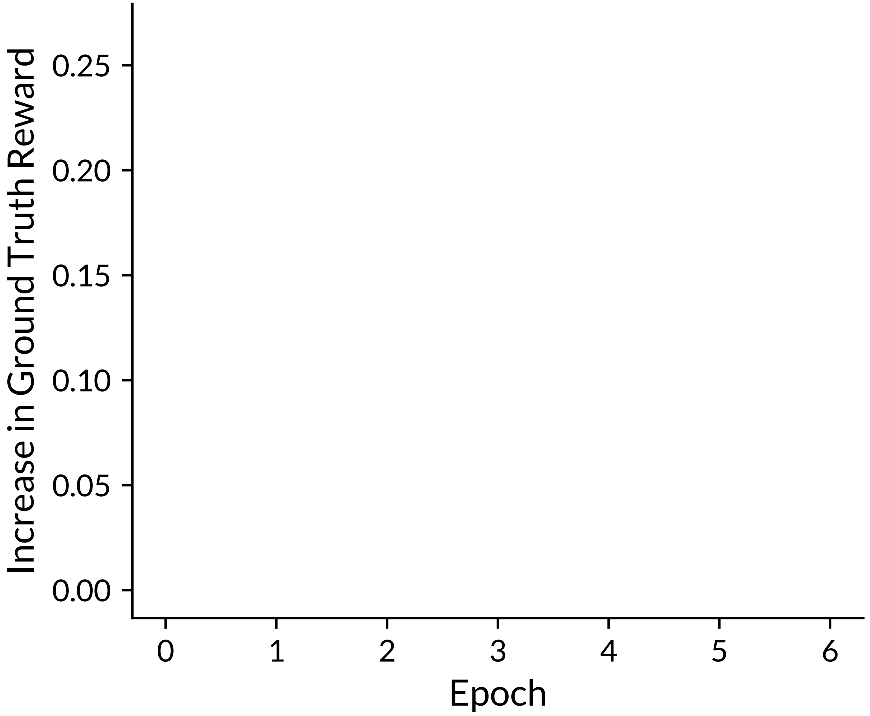
Reward Variance



Experiments: More Accurate RMs Are Not Necessarily Better

Setting:

- Ground Truth: ArmoRM
- Dataset: UltraFeedback
- LM: Pythia-2.8B

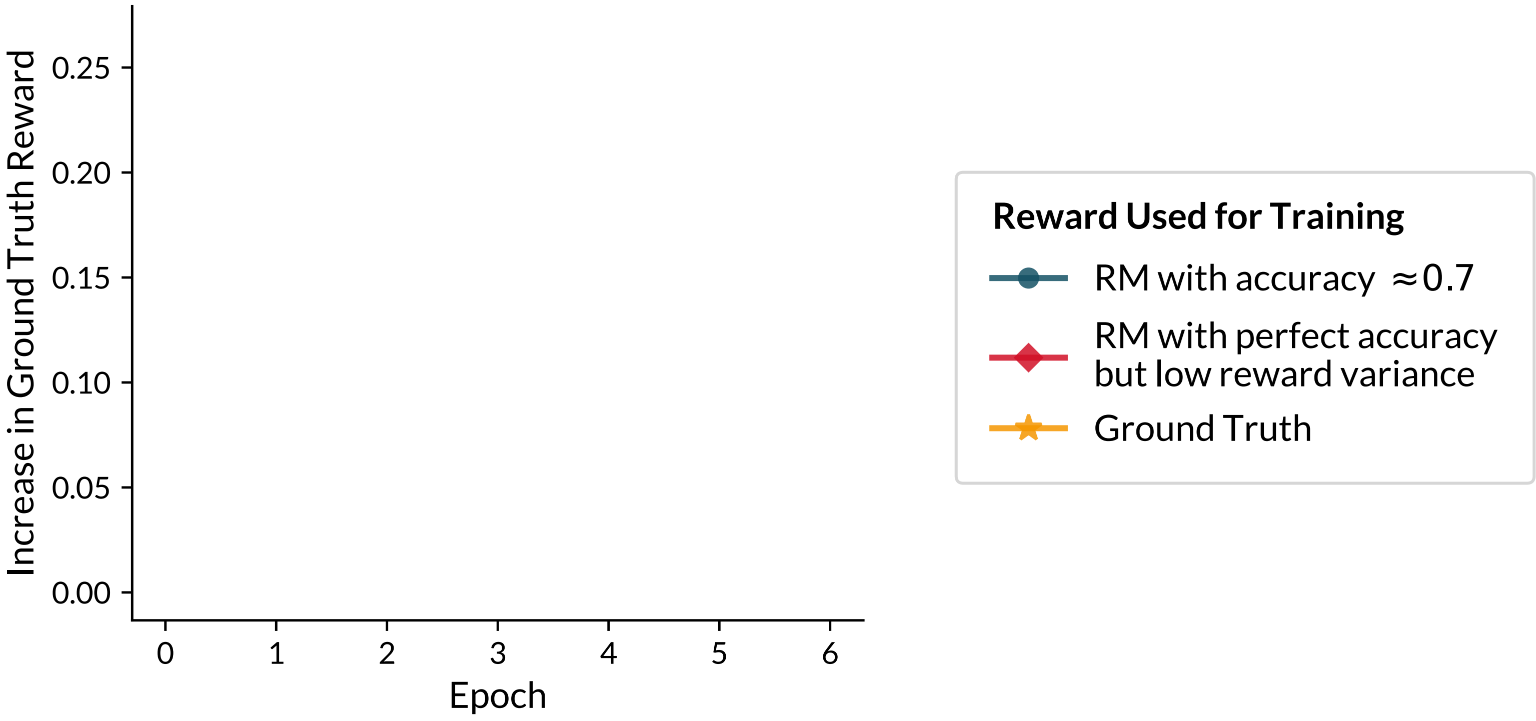


Chen et al. 2024, Wen et al. 2025: Further experiments showing more accurate RMs are not necessarily better

Experiments: More Accurate RMs Are Not Necessarily Better

Setting:

- Ground Truth: ArmoRM
- Dataset: UltraFeedback
- LM: Pythia-2.8B

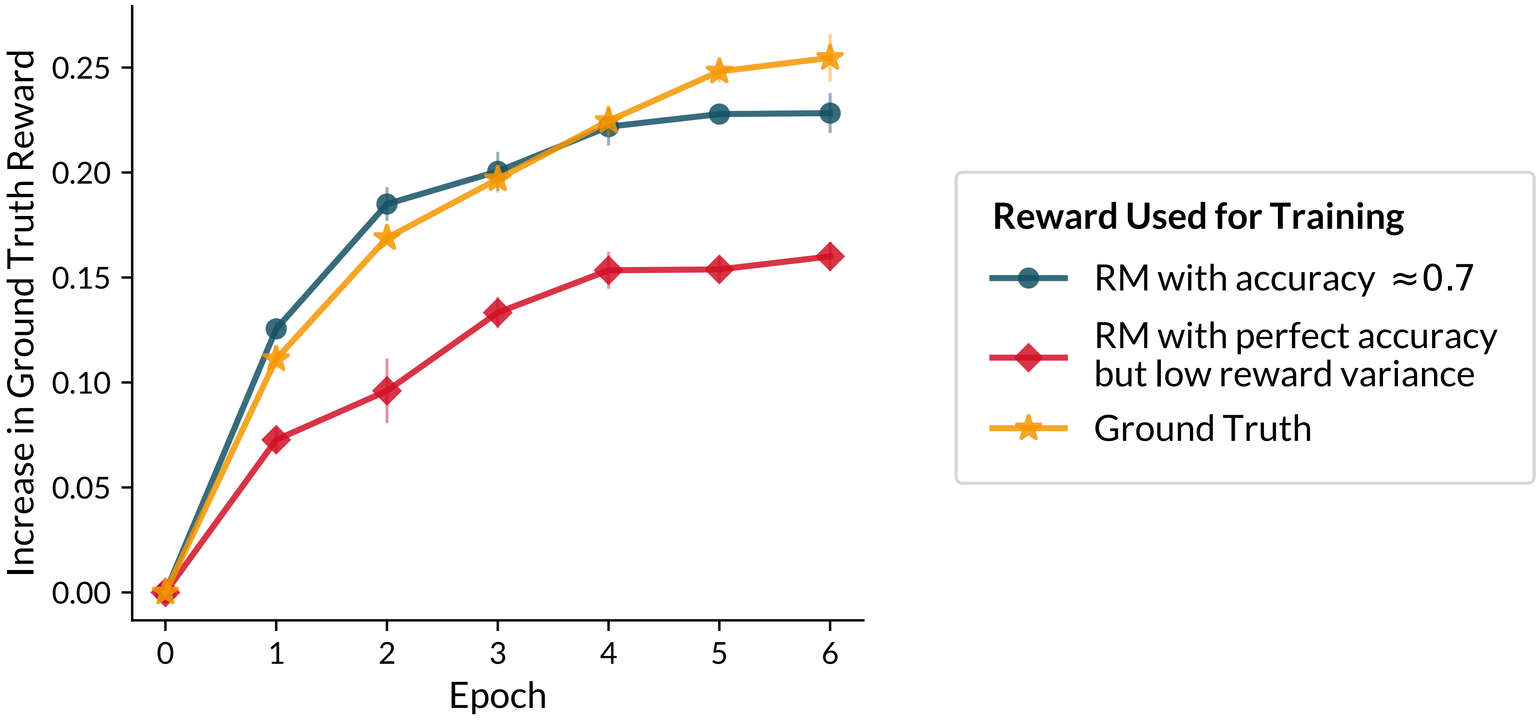


Chen et al. 2024, Wen et al. 2025: Further experiments showing more accurate RMs are not necessarily better

Experiments: More Accurate RMs Are Not Necessarily Better

Setting:

- Ground Truth: ArmoRM
- Dataset: UltraFeedback
- LM: Pythia-2.8B



Even perfectly accurate RMs can underperform less accurate ones, due to low reward variance

Chen et al. 2024, Wen et al. 2025: Further experiments showing more accurate RMs are not necessarily better

Implication II: For Different LMs, Different RMs Are Better

Observation: An RM can induce high reward variance for one LM yet low variance for another

Implication II: For Different LMs, Different RMs Are Better

Observation: An RM can induce high reward variance for one LM yet low variance for another

Theorem

There exist $r_{\mathrm{RM}}, r'_{\mathrm{RM}}$ and initial policy families Π, Π' such that:

Implication II: For Different LMs, Different RMs Are Better

Observation: An RM can induce high reward variance for one LM yet low variance for another

Theorem

There exist $r_{\mathrm{RM}}, r'_{\mathrm{RM}}$ and initial policy families Π, Π' such that:

r_{RM} is a better teacher for $\pi_{\mathrm{ref}} \in \Pi$

Implication II: For Different LMs, Different RMs Are Better

Observation: An RM can induce high reward variance for one LM yet low variance for another

Theorem

There exist $r_{\mathrm{RM}}, r'_{\mathrm{RM}}$ and initial policy families Π, Π' such that:

r_{RM} is a better teacher for $\pi_{\mathrm{ref}} \in \Pi$

r'_{RM} is a better teacher for $\pi_{\mathrm{ref}} \in \Pi'$

Implication II: For Different LMs, Different RMs Are Better

Observation: An RM can induce high reward variance for one LM yet low variance for another

Theorem

There exist $r_{\mathrm{RM}}, r'_{\mathrm{RM}}$ and initial policy families Π, Π' such that:

r_{RM} is a better teacher for $\pi_{\mathrm{ref}} \in \Pi$

r'_{RM} is a better teacher for $\pi_{\mathrm{ref}} \in \Pi'$

What makes a good RM depends on the LM being aligned

Experiments: For Different LMs, Different RMs Are Better

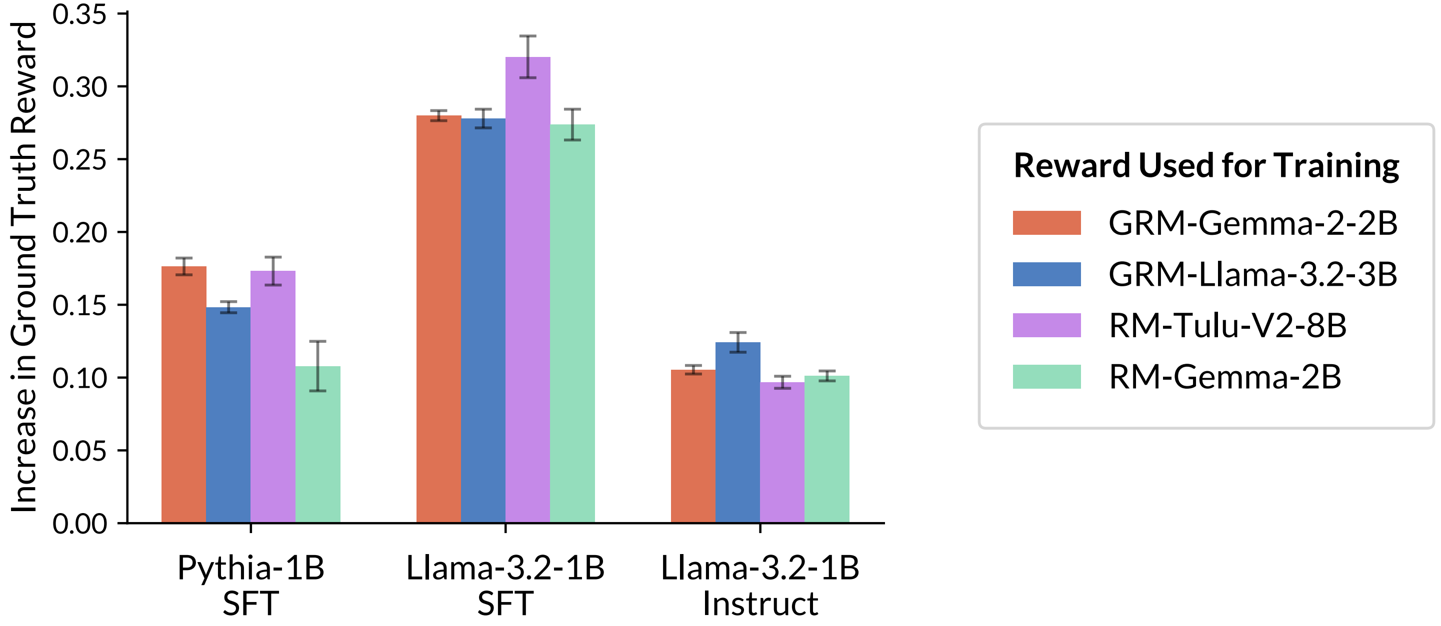
Setting:

- Ground Truth: ArmoRM
- Dataset: UltraFeedback

Experiments: For Different LMs, Different RMs Are Better

Setting:

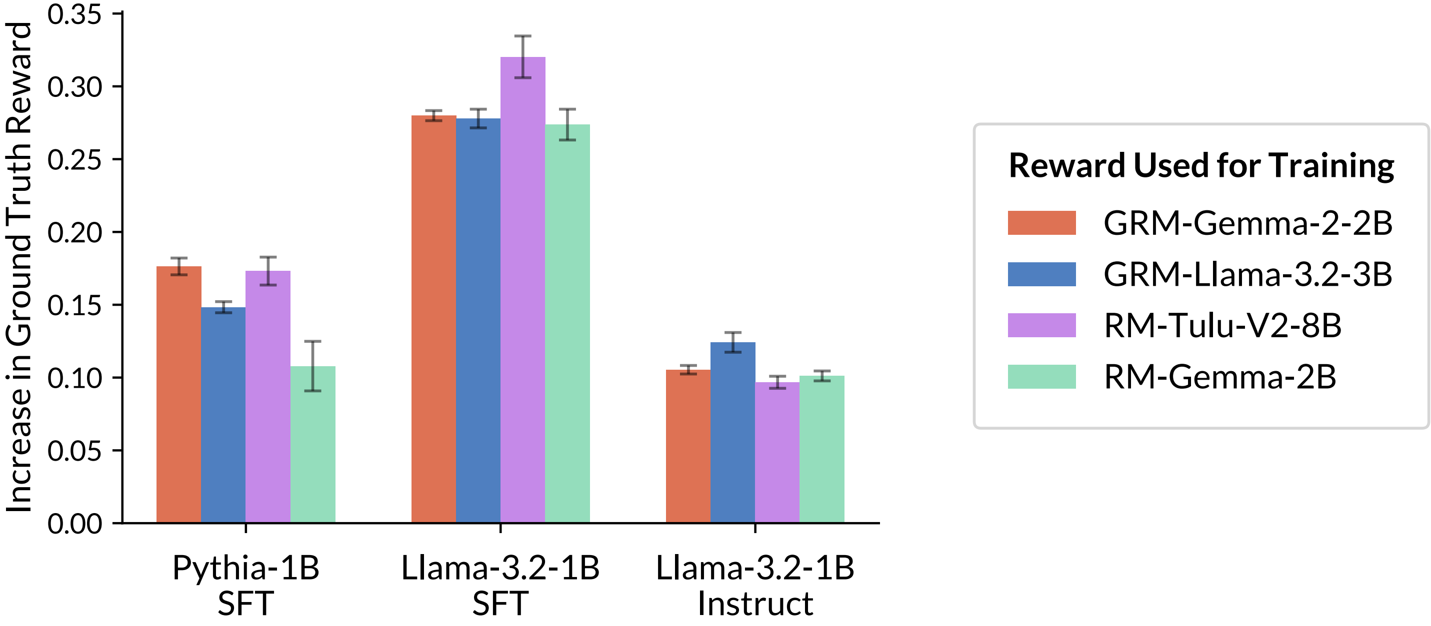
- Ground Truth: ArmoRM
- Dataset: UltraFeedback



Experiments: For Different LMs, Different RMs Are Better

Setting:

- Ground Truth: ArmoRM
- Dataset: UltraFeedback



Benchmarks evaluating RMs in isolation from the LM they guide are fundamentally limited

Takeaways: What Makes a Good RM?

Takeaways: What Makes a Good RM?

Q: What makes an RM a good teacher for RLHF?

Takeaways: What Makes a Good RM?

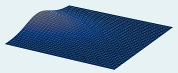
Q: What makes an RM a good teacher for RLHF?

Optimization Perspective:

Takeaways: What Makes a Good RM?

Q: What makes an RM a good teacher for RLHF?

Optimization Perspective:

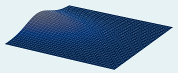


Beyond accuracy, RM needs to induce sufficient **reward variance**

Takeaways: What Makes a Good RM?

Q: What makes an RM a good teacher for RLHF?

Optimization Perspective:



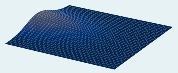
Beyond accuracy, RM needs to induce sufficient **reward variance**

More accurate RMs are not necessarily better teachers for RLHF

Takeaways: What Makes a Good RM?

Q: What makes an RM a good teacher for RLHF?

Optimization Perspective:



Beyond accuracy, RM needs to induce sufficient **reward variance**

More accurate RMs are not necessarily better teachers for RLHF

Benchmarks evaluating RMs solely based on accuracy or independently of the LM they guide are fundamentally limited

Takeaways: What Makes a Good RM?

Q: What makes an RM a good teacher for RLHF?

Optimization Perspective:

- Beyond accuracy, RM needs to induce sufficient **reward variance**
- More accurate RMs are not necessarily better teachers for RLHF
- Benchmarks evaluating RMs solely based on accuracy or independently of the LM they guide are fundamentally limited

Our results highlight the need for RM training and evaluation protocols that account for properties beyond accuracy

Importance of SFT in the RLHF Pipeline

Aside from the RM, reward variance depends on the **prompt and LM**

$$\mathrm{Var}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})\right]$$

Importance of SFT in the RLHF Pipeline

Aside from the RM, reward variance depends on the **prompt and LM**

$$\mathrm{Var}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})\right]$$

Our Results: Shed light on the importance of SFT in the RLHF pipeline

Importance of SFT in the RLHF Pipeline

Aside from the RM, reward variance depends on the **prompt and LM**

$$\mathrm{Var}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})]$$

Our Results: Shed light on the importance of SFT in the RLHF pipeline

SFT reduces number of prompts with low reward variance

Importance of SFT in the RLHF Pipeline

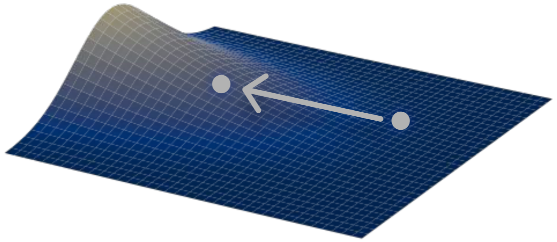
Aside from the RM, reward variance depends on the **prompt and LM**

$$\mathrm{Var}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[r_{\mathrm{RM}}(\mathbf{x},\mathbf{y})\right]$$

Our Results: Shed light on the importance of SFT in the RLHF pipeline

SFT reduces number of prompts with low reward variance

Intuition: SFT finds a less flat initialization



Practical Application I: SFT Over a Few Samples Can Suffice

Limitation of Initial SFT Phase: Requires labeled data

Practical Application I: SFT Over a Few Samples Can Suffice

Limitation of Initial SFT Phase: Requires labeled data

Our Results: Using only **1% of samples** for SFT (compared to prior work) allows RLHF to reach roughly same performance

Practical Application I: SFT Over a Few Samples Can Suffice

Limitation of Initial SFT Phase: Requires labeled data

Our Results: Using only **1% of samples** for SFT (compared to prior work) allows RLHF to reach roughly same performance

Llama 4

Kept only 5% of their SFT data for maximizing RLHF performance

Released April 5th, 2025

Practical Application II: Data Selection via Reward Variance

Data Selection Algorithms: Choose prompts for RL via reward variance

Practical Application II: Data Selection via Reward Variance

Data Selection Algorithms: Choose prompts for RL via reward variance

Not All Rollouts are Useful: Down-Sampling Rollouts in LLM Reinforcement Learning

Xu et al. 2025

Reinforcement Learning for Reasoning in Large Language Models with One Training Example

Wang et al. 2025

Learning to Reason at the Frontier of Learnability

Foster et al. 2025

Improving Generalization in Intent Detection: GRPO with Reward-Based Curriculum Sampling

Feng et al. 2025

Practical Application III: Policy Gradient Methods

∇ **Policy Gradient Methods:** Develop new update rules and reward transformations

Practical Application III: Policy Gradient Methods

∇ **Policy Gradient Methods:** Develop new update rules and reward transformations

Accelerating RLHF Training with Reward Variance Increase
Yang et al. 2025

DGRO: Enhancing LLM Reasoning via Exploration-Exploitation Control and Reward Variance Management
Su et al. 2025

RePO: Replay-Enhanced Policy Optimization
Li et al. 2025

ReDit: Reward Dithering for Improved LLM Policy Optimization
Wei et al. 2025

Takeaways: Importance of Reward Variance

Reward variance is a key quantity for successful RLHF

Takeaways: Importance of Reward Variance

Reward variance is a key quantity for successful RLHF

Can help identify optimization issues

Takeaways: Importance of Reward Variance

Reward variance is a key quantity for successful RLHF

- Can help identify optimization issues
- Useful for developing data selection, policy gradient, and RM training algorithms

Difference Between RM Types

We Discussed: How properties of RM affect RLHF

Difference Between RM Types

We Discussed: How $\underbrace{\text{properties of RM}}_{\text{depend on RM type}}$ affect RLHF

Difference Between RM Types

We Discussed: How $\underbrace{\text{properties of RM}}_{\text{depend on RM type}}$ affect RLHF

What are the pros and cons of different types?

Difference Between RM Types

We Discussed: How properties of RM affect RLHF

depend on **RM type**

What are the pros and cons of different types?

Why is Your Language Model a Poor Implicit Reward Model?

Noam Razin[†], Yong Lin[†], Jiarui Yao[‡], Sanjeev Arora[†]

[†] Princeton Language and Intelligence, Princeton University
[‡] University of Illinois Urbana-Champaign

Difference Between RM Types

We Discussed: How properties of RM affect RLHF

depend on **RM type**

What are the pros and cons of different types?

Why is Your Language Model a Poor Implicit Reward Model?

Noam Razin†, Yong Lin†, Jiarui Yao‡, Sanjeev Arora†

† Princeton Language and Intelligence, Princeton University
‡ University of Illinois Urbana-Champaign

Our Results: Reveal why RM types generalize differently (in terms of accuracy)

Difference Between RM Types

We Discussed: How $\underbrace{\text{properties of RM}}_{\text{depend on RM type}}$ affect RLHF

depend on **RM type**

What are the pros and cons of different types?

Why is Your Language Model a Poor Implicit Reward Model?

Noam Razin[†], **Yong Lin**[†], **Jiarui Yao**[‡], **Sanjeev Arora**[†]

[†] Princeton Language and Intelligence, Princeton University
[‡] University of Illinois Urbana-Champaign

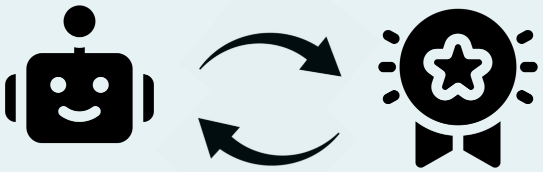
Our Results: Reveal why RM types generalize differently (in terms of accuracy)

Seemingly minor design choices can substantially affect reward model generalization

Part II: Alignment via Direct Preference Learning

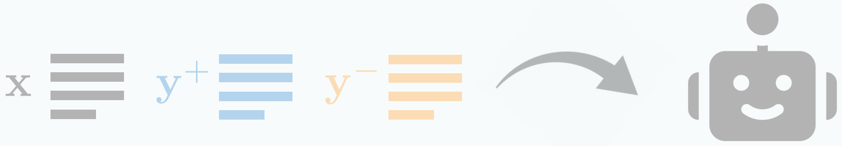
Reinforcement Learning

(e.g. Ouyang et al. 2022)



Direct Preference Learning

(e.g. Rafailov et al. 2023)



Vanishing Gradients in Reinforcement Finetuning of Language Models

R + Zhou + Saremi + Thilak + Bradley + Nakkiran + Susskind + Littwin | *ICLR 2024*

What Makes a Reward Model a Good Teacher? An Optimization Perspective

R + Wang + Strauss + Wei + Lee + Arora | *arXiv 2025*

Why is Your Language Model a Poor Implicit Reward Model?

R + Lin + Yao + Arora | *arXiv 2025*

Part II: Alignment via Direct Preference Learning

Reinforcement Learning

(e.g. Ouyang et al. 2022)

Direct Preference Learning

(e.g. Rafailov et al. 2023)

Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization

R + Malladi + Bhaskar + Chen + Arora + Hanin |
ICLR 2025

Finetuning LMs via Direct Preference Learning

RLHF can be computationally expensive and unstable

Rafailov et al. 2023, Azar et al. 2024, Tang et al. 2024, Xu et al. 2024, Meng et al. 2024

Finetuning LMs via Direct Preference Learning

RLHF can be computationally expensive and unstable

Direct Preference Learning

Directly train the LM over the preference data (e.g. DPO)

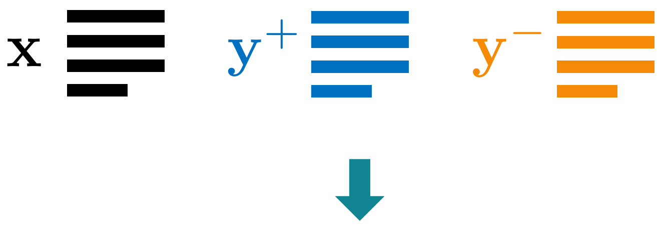
Rafailov et al. 2023, Azar et al. 2024, Tang et al. 2024, Xu et al. 2024, Meng et al. 2024

Finetuning LMs via Direct Preference Learning

RLHF can be computationally expensive and unstable

Direct Preference Learning

Directly train the LM over the preference data (e.g. DPO)



$$\ell\big(\ln \pi_\theta(\mathbf{y}^+|\mathbf{x}) - \ln \pi_\theta(\mathbf{y}^-|\mathbf{x})\big)$$

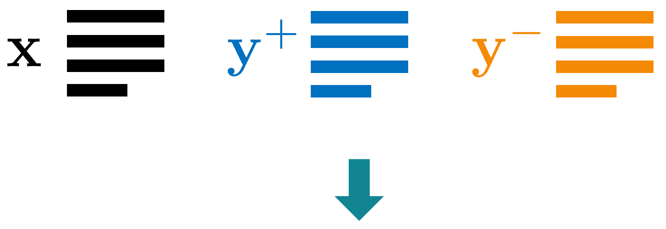
Rafailov et al. 2023, Azar et al. 2024, Tang et al. 2024, Xu et al. 2024, Meng et al. 2024

Finetuning LMs via Direct Preference Learning

RLHF can be computationally expensive and unstable

Direct Preference Learning

Directly train the LM over the preference data (e.g. DPO)



$$\ell\Big(\ln \pi_\theta\left(\mathbf{y}^+ | \mathbf{x}\right) - \ln \pi_\theta\left(\mathbf{y}^- | \mathbf{x}\right) \Big)$$

Numerous variants of DPO, differing in choice of ℓ

Rafailov et al. 2023, Azar et al. 2024, Tang et al. 2024, Xu et al. 2024, Meng et al. 2024

Finetuning LMs via Direct Preference Learning

RLHF can be computationally expensive and unstable

Direct Preference Learning

Directly train the LM over the preference data (e.g. DPO)

$\mathbf{x}$ $\quad$ $\mathbf{y}^+$ $\quad$ $\mathbf{y}^-$

$$\ell\Big(\ln \pi_\theta\left(\mathbf{y}^+|\mathbf{x}\right) - \ln \pi_\theta\left(\mathbf{y}^-|\mathbf{x}\right) \Big)$$

Numerous variants of DPO, differing in choice of ℓ

Intuitively, $\pi_\theta\left(\mathbf{y}^+|\mathbf{x}\right)$ should increase and $\pi_\theta\left(\mathbf{y}^-|\mathbf{x}\right)$ should decrease

Rafailov et al. 2023, Azar et al. 2024, Tang et al. 2024, Xu et al. 2024, Meng et al. 2024

Likelihood Displacement

However, the probability of preferred responses **often decreases!**

(Pal et al. 2024; Yuan et al. 2024, Rafailov et al. 2024, Tajwar et al. 2024, Pang et al. 2024, Liu et al. 2024)

Likelihood Displacement

However, the probability of preferred responses **often decreases!**

(Pal et al. 2024; Yuan et al. 2024, Rafailov et al. 2024, Tajwar et al. 2024, Pang et al. 2024, Liu et al. 2024)

Likelihood Displacement

Likelihood Displacement

However, the probability of preferred responses **often decreases!**

(Pal et al. 2024; Yuan et al. 2024, Rafailov et al. 2024, Tajwar et al. 2024, Pang et al. 2024, Liu et al. 2024)

Likelihood Displacement

Benign

$\mathbf{z}$ is similar in meaning to $\mathbf{y}^+$

Catastrophic

$\mathbf{z}$ is opposite in meaning to $\mathbf{y}^+$

Likelihood Displacement

However, the probability of preferred responses **often decreases!**

(Pal et al. 2024; Yuan et al. 2024, Rafailov et al. 2024, Tajwar et al. 2024, Pang et al. 2024, Liu et al. 2024)

Likelihood Displacement

Benign
$\mathbf{z}$ is similar in meaning to $\mathbf{y}^+$

Catastrophic
$\mathbf{z}$ is opposite in meaning to $\mathbf{y}^+$

Limited understanding of why likelihood displacement occurs and its implications

Likelihood Displacement Can Cause Unintentional Unalignment

Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a language model to refuse unsafe prompts via DPO

Likelihood Displacement Can Cause Unintentional Unalignment

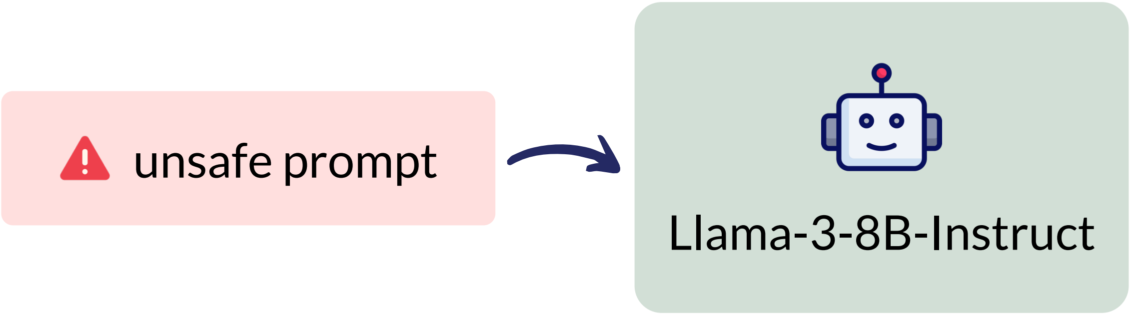
Setting: Train a language model to refuse unsafe prompts via DPO

Preference Dataset: Unsafe prompts from SORRY-Bench (Xie et al. 2024)

Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a language model to refuse unsafe prompts via DPO

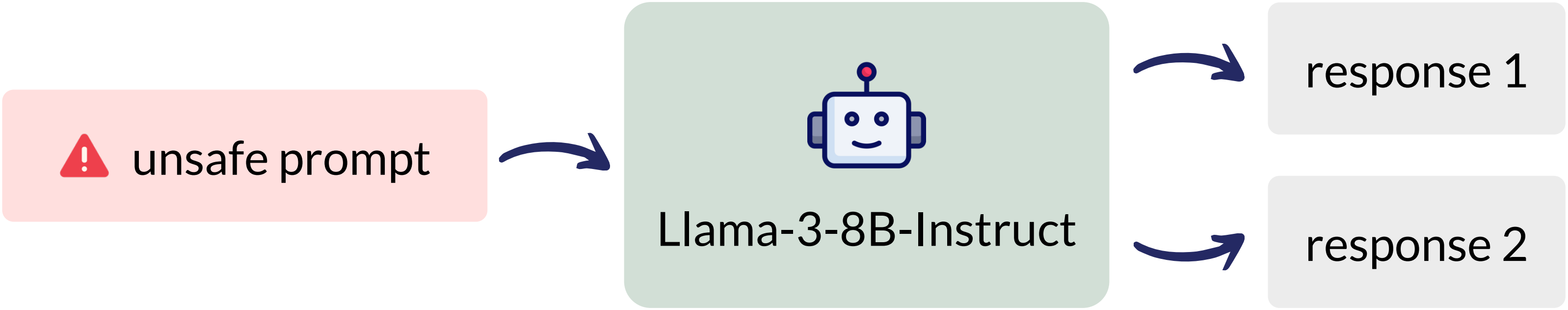
Preference Dataset: Unsafe prompts from SORRY-Bench (Xie et al. 2024)



Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a language model to refuse unsafe prompts via DPO

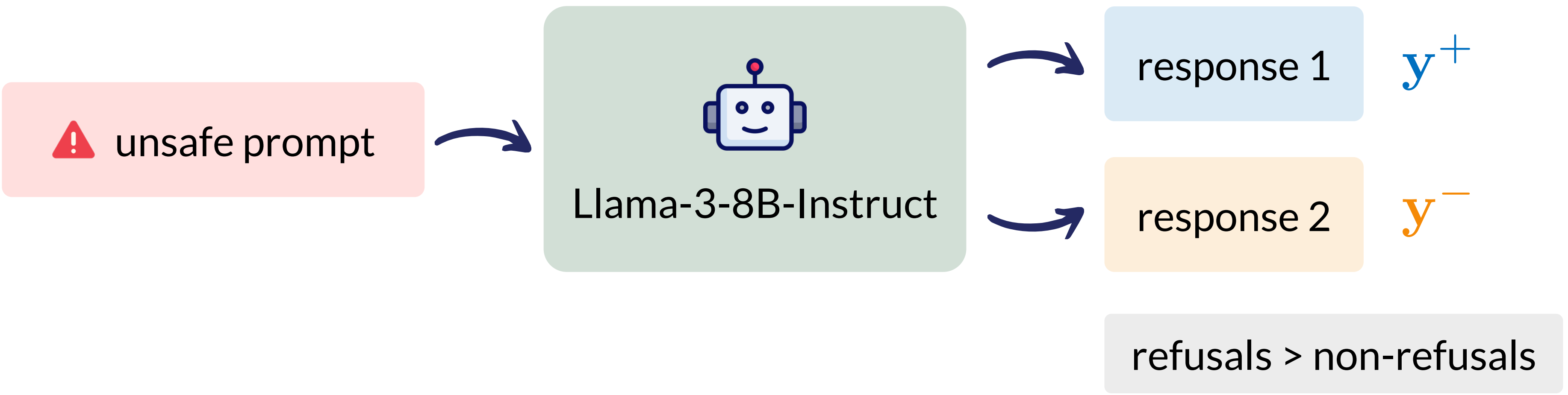
Preference Dataset: Unsafe prompts from SORRY-Bench (Xie et al. 2024)



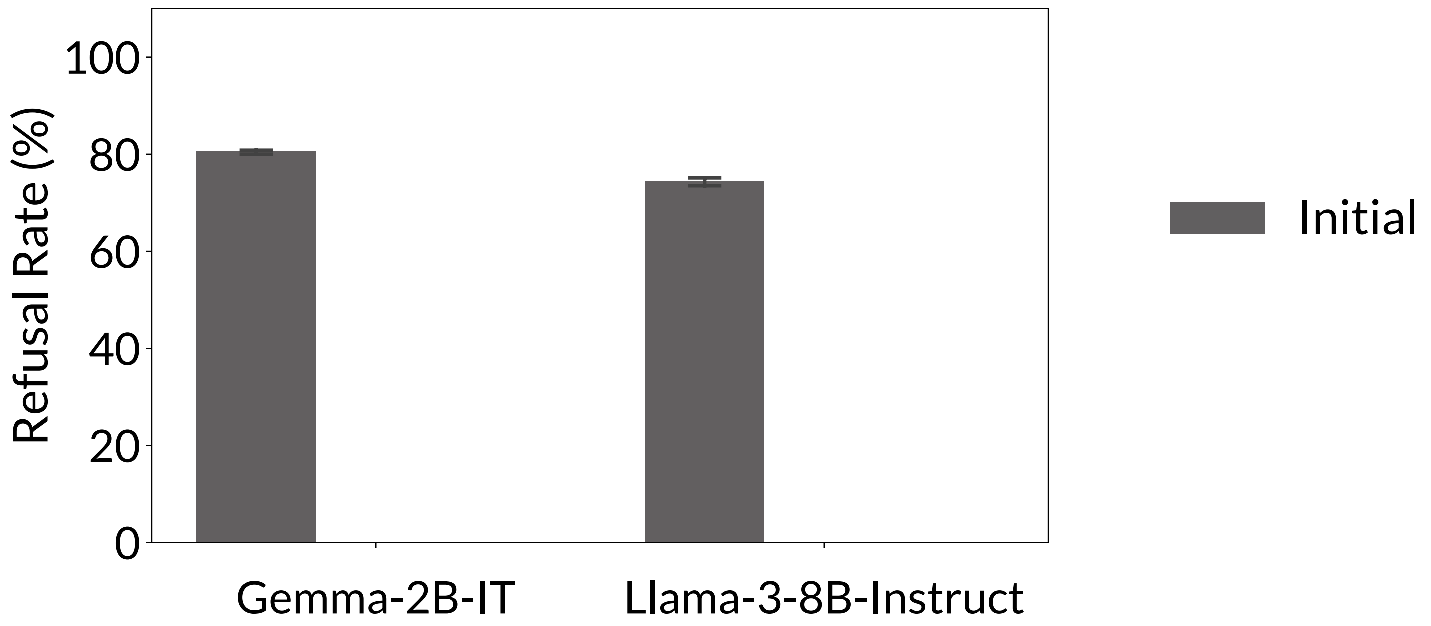
Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a language model to refuse unsafe prompts via DPO

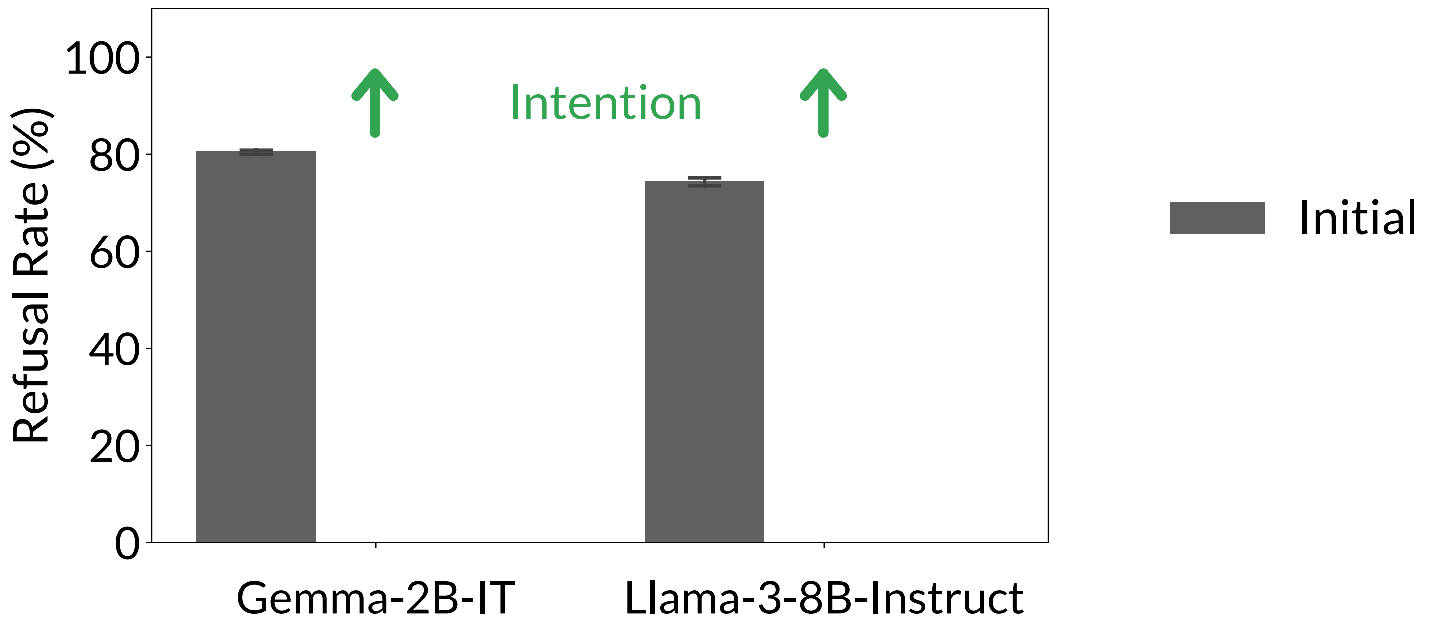
Preference Dataset: Unsafe prompts from SORRY-Bench (Xie et al. 2024)



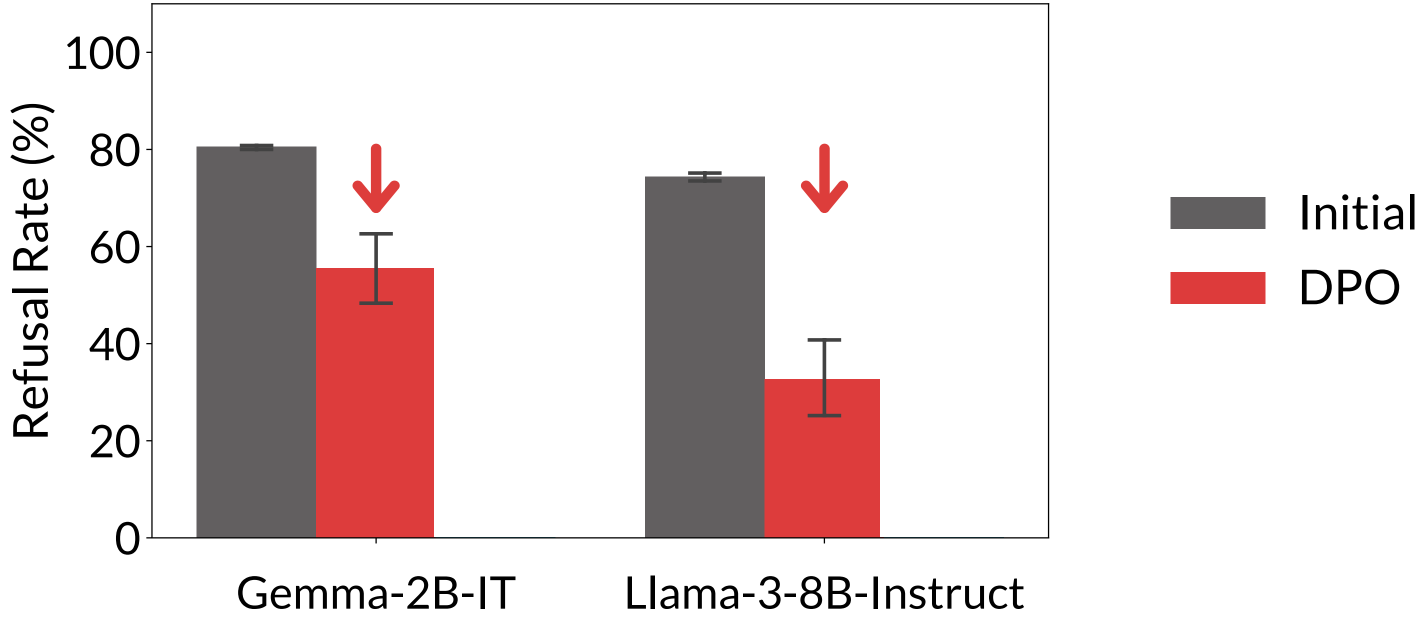
Likelihood Displacement Can Cause Unintentional Unalignment



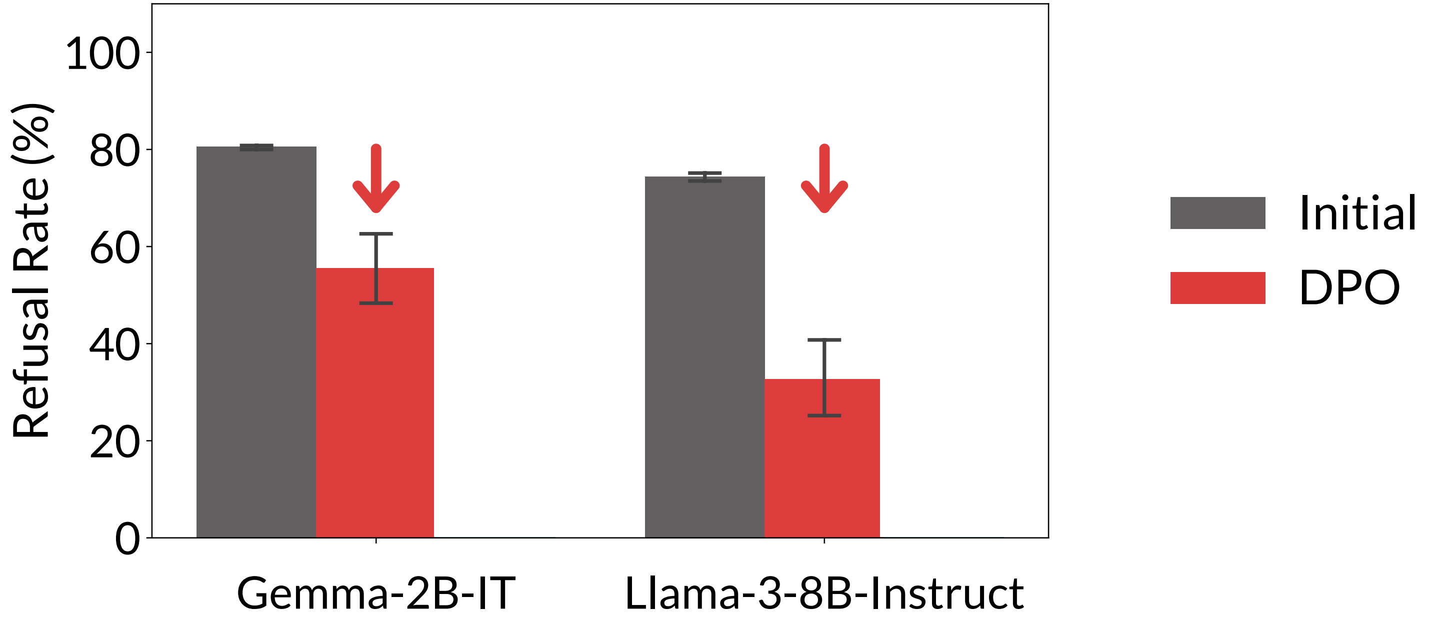
Likelihood Displacement Can Cause Unintentional Unalignment



Likelihood Displacement Can Cause Unintentional Unalignment



Likelihood Displacement Can Cause Unintentional Unalignment



Likelihood displacement leads to unintentional unalignment!

Theoretical Analysis of Likelihood Displacement

Theoretical Analysis of Likelihood Displacement

Approach: Characterize how $\pi_\theta(\mathbf{y}^+|\mathbf{x})$ changes during training

Theoretical Analysis of Likelihood Displacement

Approach: Characterize how $\pi_\theta(\mathbf{y}^+|\mathbf{x})$ changes during training

Our Theory: Preferences with **similar hidden embeddings** lead to likelihood displacement

Theoretical Analysis of Likelihood Displacement

Approach: Characterize how $\pi_\theta(\mathbf{y}^+|\mathbf{x})$ changes during training

Our Theory: Preferences with **similar hidden embeddings** lead to likelihood displacement

Definition: Centered Hidden Embedding Similarity (CHES) Score

$$\mathrm{CHES}_{\mathbf{x}}(\mathbf{y}^+, \mathbf{y}^-) :=$$

Theoretical Analysis of Likelihood Displacement

Approach: Characterize how $\pi_\theta(\mathbf{y}^+|\mathbf{x})$ changes during training

Our Theory: Preferences with **similar hidden embeddings** lead to likelihood displacement

Definition: Centered Hidden Embedding Similarity (CHES) Score

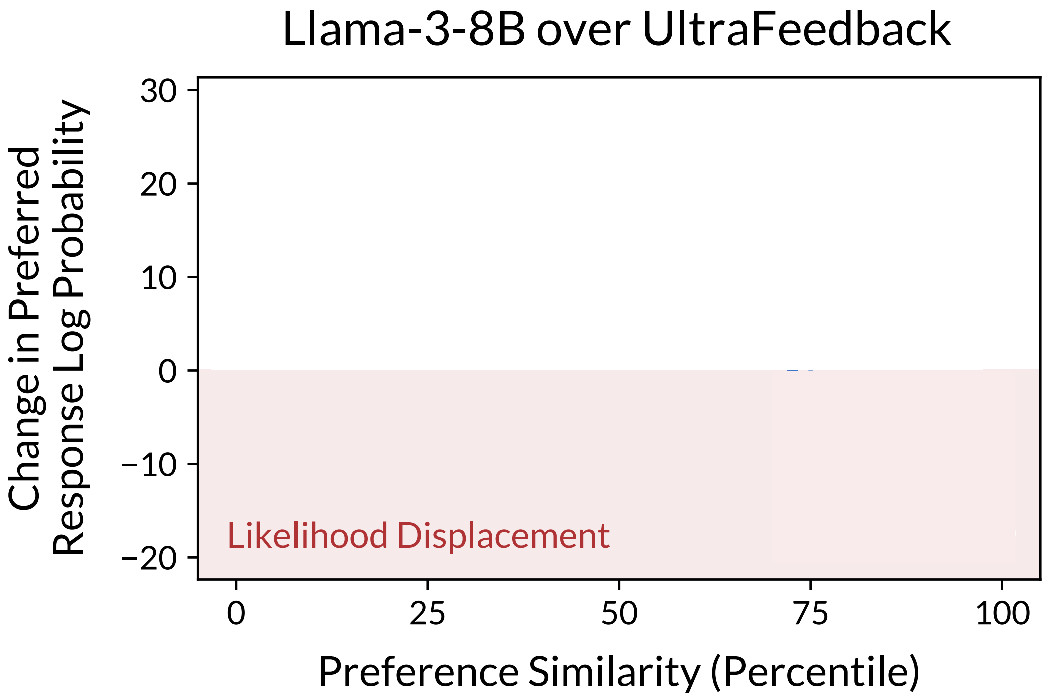
$$\mathrm{CHES}_{\mathbf{x}}(\mathbf{y}^+, \mathbf{y}^-) := \Big\langle \underbrace{\sum_{k=1}^{|\mathbf{y}^+|} \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k}}}_{\mathbf{y}^+ \text{ embeddings}}, \underbrace{\sum_{k'=1}^{|\mathbf{y}^-|} \mathbf{h}_{\mathbf{x},\mathbf{y}^-_{<k'}}}_{\mathbf{y}^- \text{ embeddings}} \Big\rangle - \Big\| \sum_{k=1}^{|\mathbf{y}^+|} \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k}} \Big\|^2$$

Identifying Sources of Likelihood Displacement

Q: How indicative is the CHES score of likelihood displacement?

Identifying Sources of Likelihood Displacement

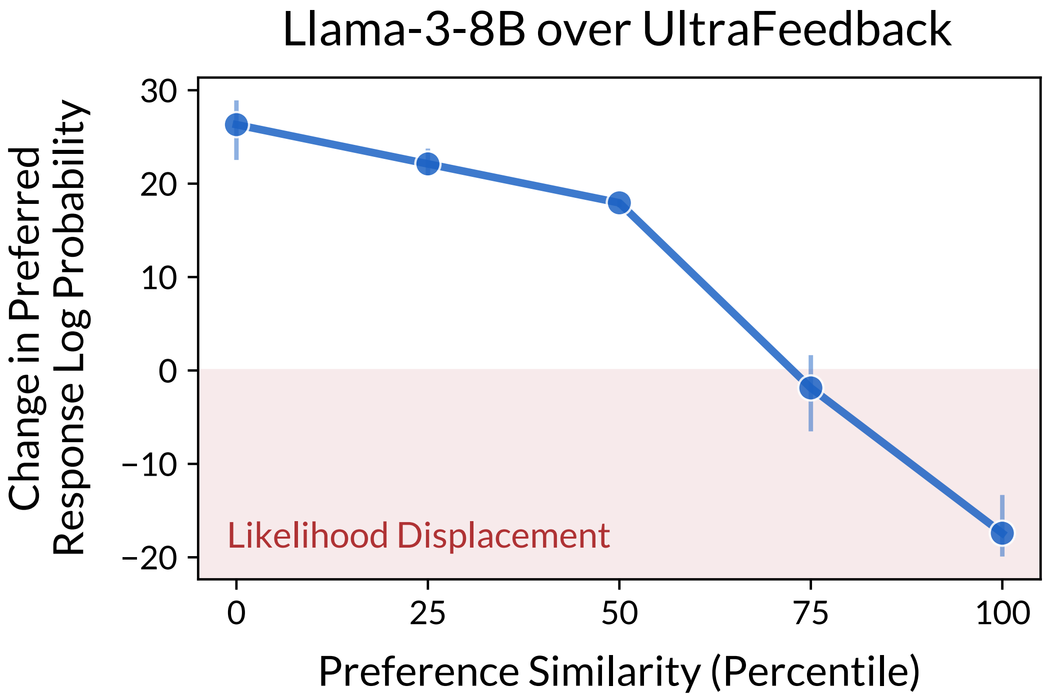
Q: How indicative is the CHES score of likelihood displacement?



*Similar results for OLMo-1B, Gemma-2B models and AlpacaFarm dataset

Identifying Sources of Likelihood Displacement

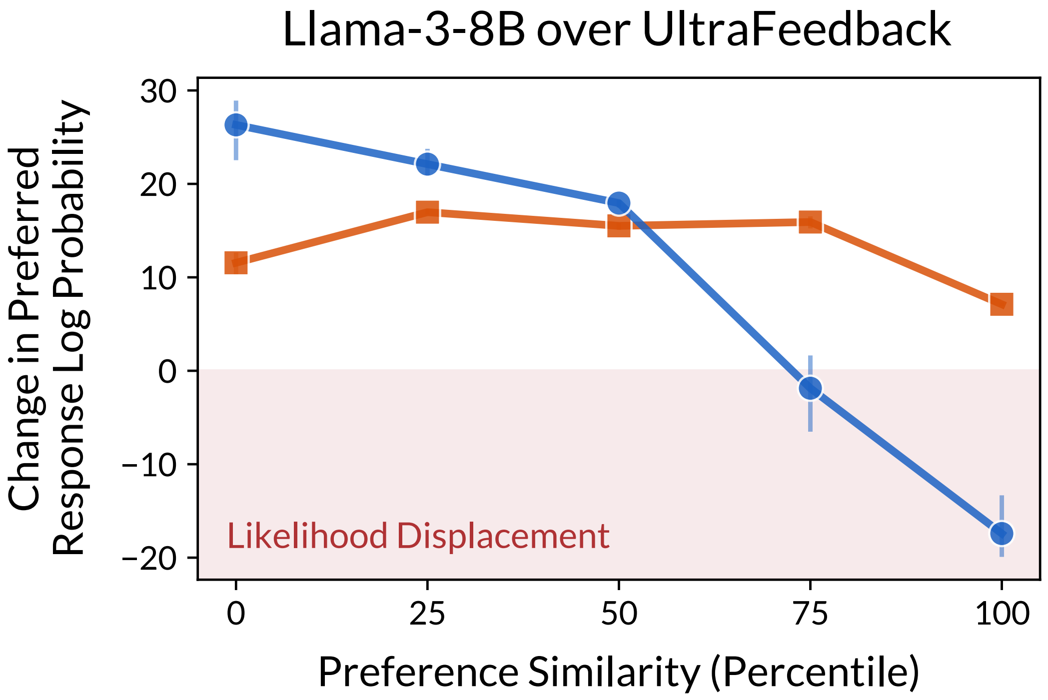
Q: How indicative is the CHES score of likelihood displacement?



*Similar results for OLMo-1B, Gemma-2B models and AlpacaFarm dataset

Identifying Sources of Likelihood Displacement

Q: How indicative is the CHES score of likelihood displacement?



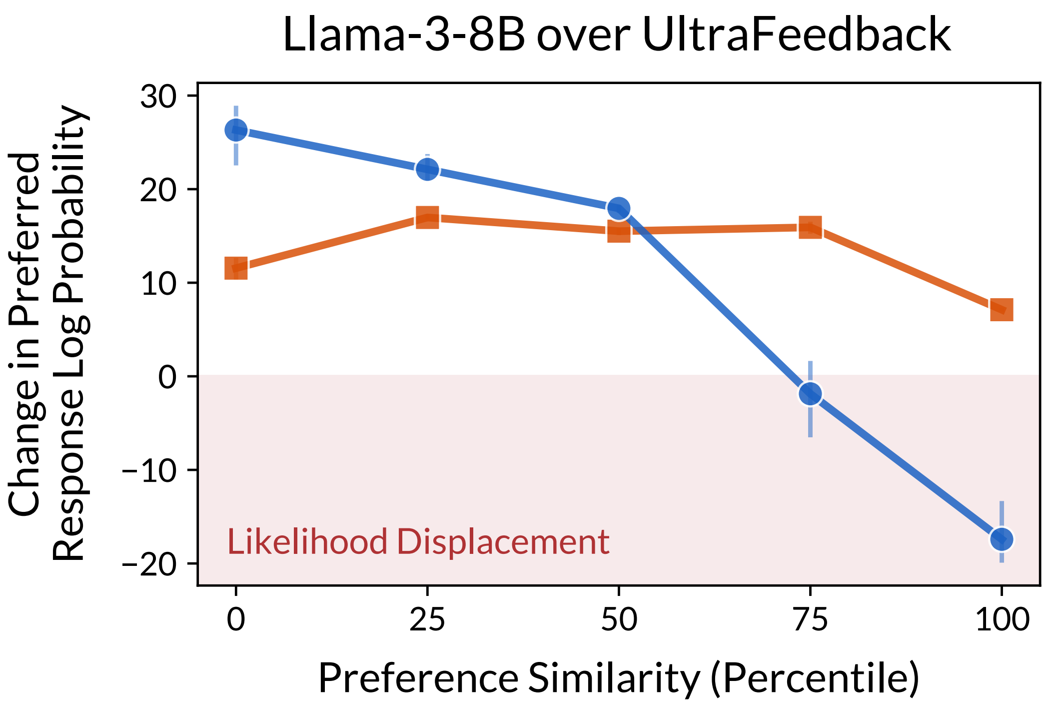
*Similar results for OLMo-1B, Gemma-2B models and AlpacaFarm dataset

CHES Score

Edit Distance Similarity (Pal et al. 2024)

Identifying Sources of Likelihood Displacement

Q: How indicative is the CHES score of likelihood displacement?



*Similar results for OLMo-1B, Gemma-2B models and AlpacaFarm dataset

CHES Score

Edit Distance Similarity (Pal et al. 2024)

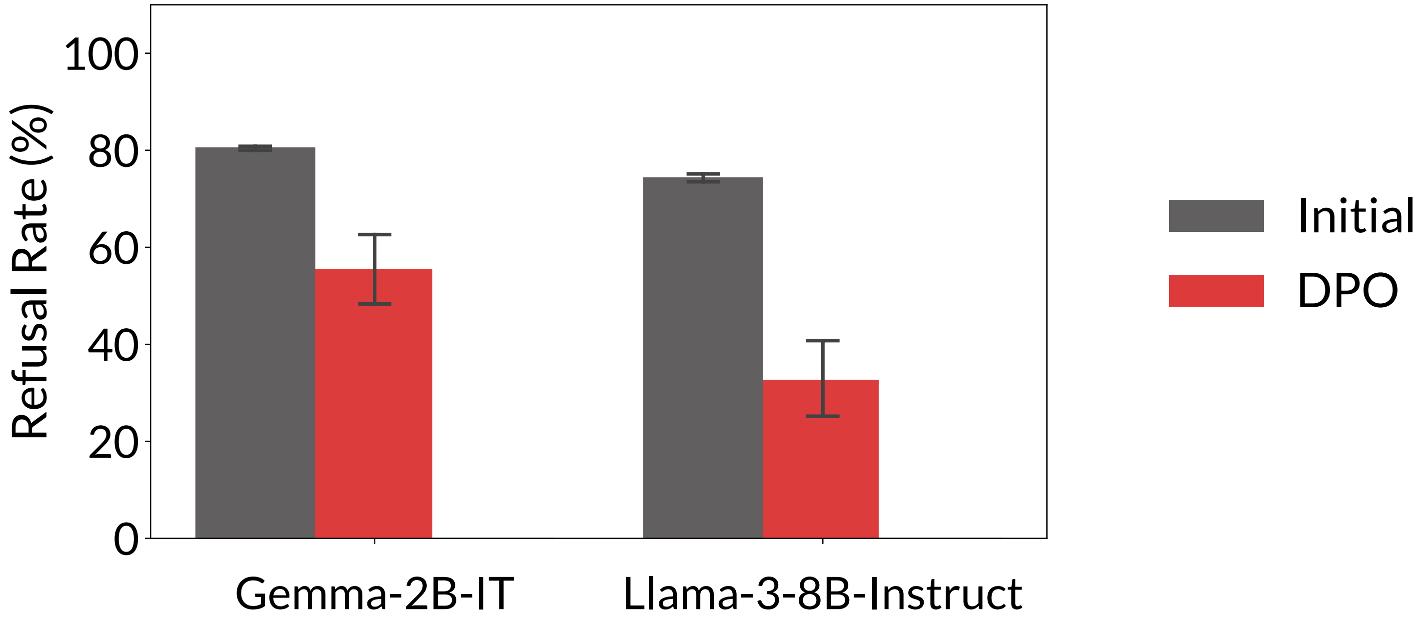
CHES score identifies training samples causing likelihood displacement, whereas alternative measures do not

Mitigating Unintentional Unalignment via Data Filtering

Recall: Unintentional unalignment due to likelihood displacement experiments

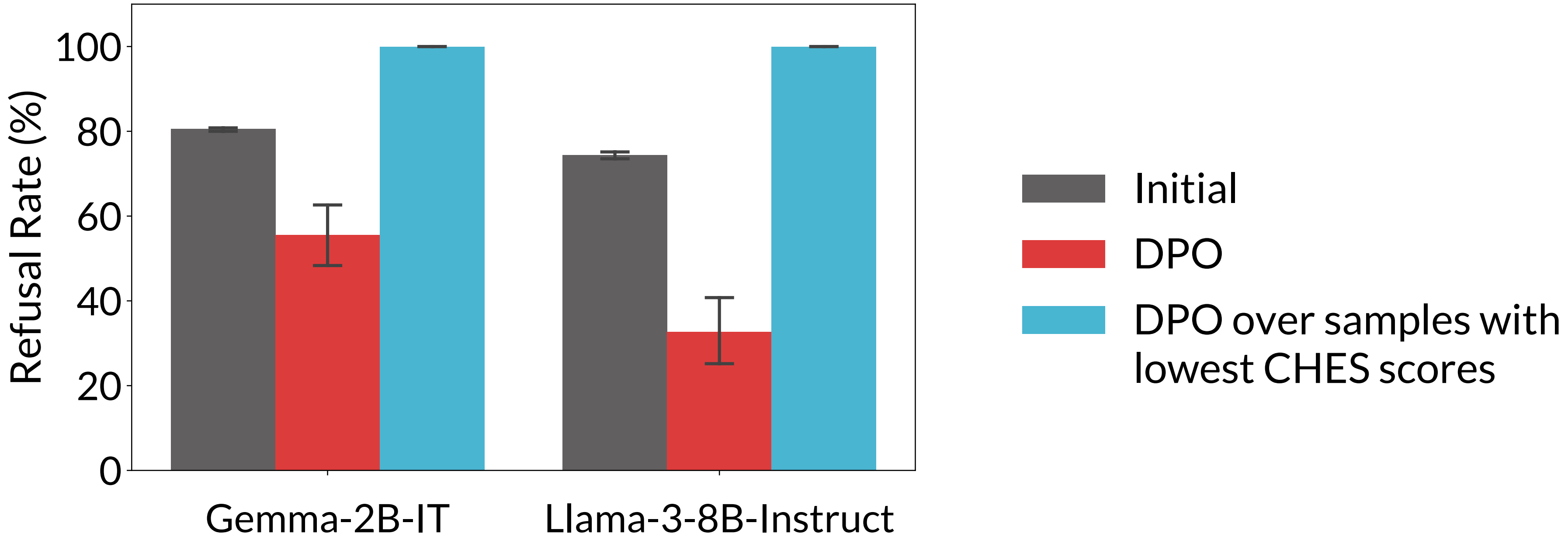
Mitigating Unintentional Unalignment via Data Filtering

Recall: Unintentional unalignment due to likelihood displacement experiments



Mitigating Unintentional Unalignment via Data Filtering

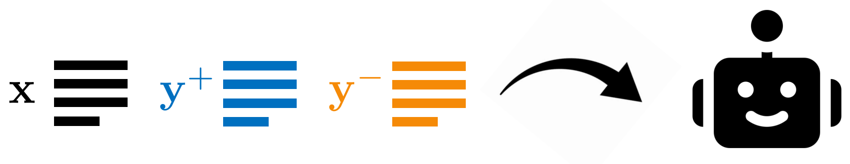
Recall: Unintentional unalignment due to likelihood displacement experiments



Removing samples with high CHES scores mitigates unintentional unalignment

Practical Impact

Our work inspired **new direct preference learning algorithms** for mitigating likelihood displacement



Practical Impact

Our work inspired **new direct preference learning algorithms** for mitigating likelihood displacement

ComPO: Preference Alignment via Comparison Oracles

Chen et al. 2025

AlphaPO: Reward Shape Matters for LLM Alignment

Gupta et al. 2025

DPO-Shift: Shifting the Distribution of Direct Preference Optimization

Yang et al. 2025

Decoupling Contrastive Decoding: Robust Hallucination Mitigation in Multimodal Large Language Models

Chen et al. 2025

Conclusion

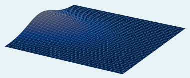
Recap

Reinforcement Learning (RLHF)

Direct Preference Learning

Recap

Reinforcement Learning (RLHF)

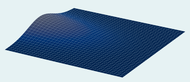


Beyond accuracy, RM needs to induce sufficient **reward variance**

Direct Preference Learning

Recap

Reinforcement Learning (RLHF)



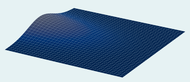
Beyond accuracy, RM needs to induce sufficient **reward variance**

Implications: More accurate RMs are not better teachers for RLHF + existing RM benchmarks are fundamentally limited

Direct Preference Learning

Recap

Reinforcement Learning (RLHF)



Beyond accuracy, RM needs to induce sufficient **reward variance**

Implications: More accurate RMs are not better teachers for RLHF + existing RM benchmarks are fundamentally limited

Practical Applications: Data selection and policy gradient methods

Direct Preference Learning

Recap

Reinforcement Learning (RLHF)

Beyond accuracy, RM needs to induce sufficient **reward variance**

Implications: More accurate RMs are not better teachers for RLHF + existing RM benchmarks are fundamentally limited

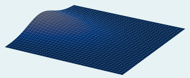
Practical Applications: Data selection and policy gradient methods

Direct Preference Learning

Likelihood displacement can cause **unintentional unalignment**

Recap

Reinforcement Learning (RLHF)



Beyond accuracy, RM needs to induce sufficient **reward variance**

Implications: More accurate RMs are not better teachers for RLHF + existing RM benchmarks are fundamentally limited

Practical Applications: Data selection and policy gradient methods

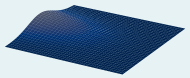
Direct Preference Learning

Likelihood displacement can cause **unintentional unalignment**

Theory & Experiments: Samples with **high CHES scores lead to likelihood displacement**

Recap

Reinforcement Learning (RLHF)



Beyond accuracy, RM needs to induce sufficient **reward variance**

Implications: More accurate RMs are not better teachers for RLHF + existing RM benchmarks are fundamentally limited

Practical Applications: Data selection and policy gradient methods

Direct Preference Learning

Likelihood displacement can cause **unintentional unalignment**

Theory & Experiments: Samples with **high CHES scores lead to likelihood displacement**

Practical Applications: Data curation and direct preference learning algorithms

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF	DPO	IPO	SimPO	KTO	...
Ouyang et al. 2022	Rafailov et al. 2023	Azar et al. 2023	Meng et al. 2024	Ethayarajh et al. 2024	

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF	DPO	IPO	SimPO	KTO	...
Ouyang et al. 2022	Rafailov et al. 2023	Azar et al. 2023	Meng et al. 2024	Ethayarajh et al. 2024	

As We Saw: Limited understanding can lead to undesirable outcomes

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF Ouyang et al. 2022	DPO Rafailov et al. 2023	IPO Azar et al. 2023	SimPO Meng et al. 2024	KTO Ethayarajh et al. 2024	...

As We Saw: Limited understanding can lead to undesirable outcomes

- Inefficient training
- Safety concerns

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF	DPO	IPO	SimPO	KTO	...
Ouyang et al. 2022	Rafailov et al. 2023	Azar et al. 2023	Meng et al. 2024	Ethayarajh et al. 2024	

As We Saw: Limited understanding can lead to undesirable outcomes

- Inefficient training
- Safety concerns

Mistakes are costly due to the large scale of current models

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF	DPO	IPO	SimPO	KTO	...
Ouyang et al. 2022	Rafailov et al. 2023	Azar et al. 2023	Meng et al. 2024	Ethayarajh et al. 2024	

As We Saw: Limited understanding can lead to undesirable outcomes

- Inefficient training
- Safety concerns

Mistakes are costly due to the large scale of current models

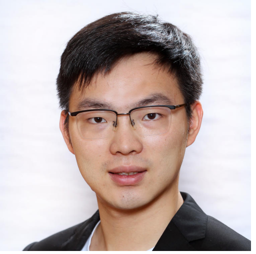
Theory (mathematical or empirical) may be necessary for efficient and reliable deployment of modern AI systems

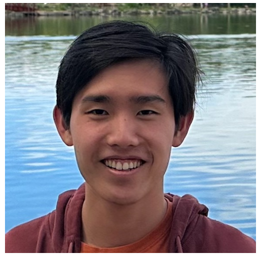














Thank You!

Work supported in part by the
Zuckerman STEM Leadership Program