Vanishing Gradients in Reinforcement Finetuning of Language Models

Noam Razin

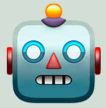
Joint work with Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley,
Preetum Nakkiran, Joshua Susskind, Etai Littwin

DLCT 9 February 2024

Language Models (LMs)

Language Models (LMs)

LM – Neural network trained on large amounts of (internet) text data
to produce a **distribution over text**

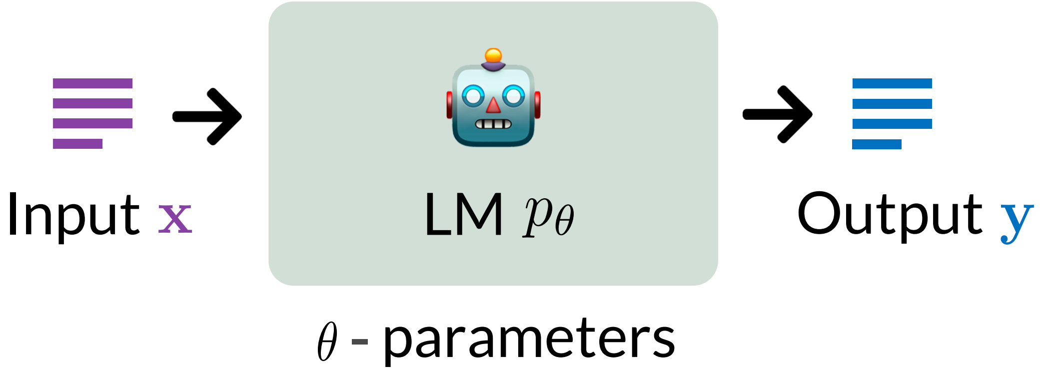


LM p_θ

θ - parameters

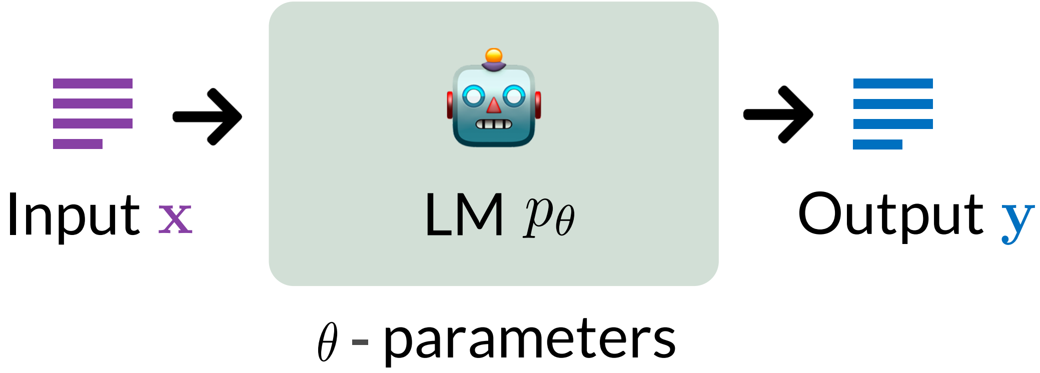
Language Models (LMs)

LM – Neural network trained on large amounts of (internet) text data to produce a **distribution over text**



Language Models (LMs)

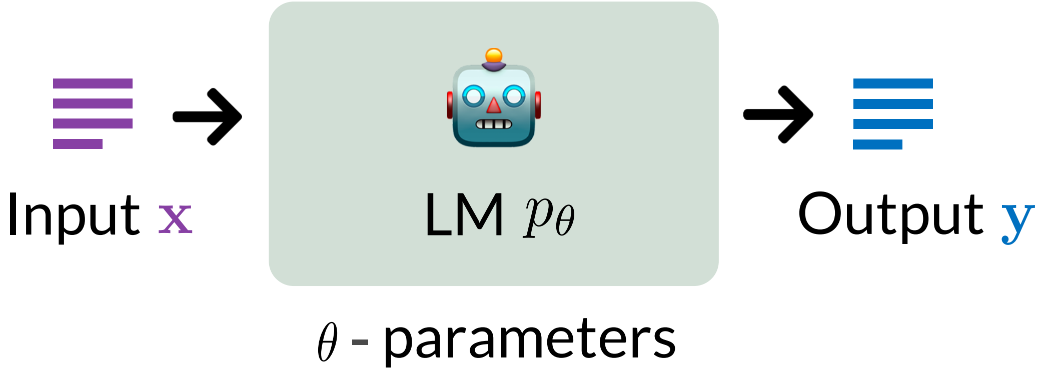
LM – Neural network trained on large amounts of (internet) text data to produce a **distribution over text**



LMs are typically **autoregressive:** $p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{l=1}^{L} p_\theta(y_l|\mathbf{x}, \mathbf{y}_{\leq l-1})$

Language Models (LMs)

LM – Neural network trained on large amounts of (internet) text data to produce a **distribution over text**



LMs are typically **autoregressive:** $p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{l=1}^{L} \underbrace{p_\theta(y_l|\mathbf{x}, \mathbf{y}_{\leq l-1})}$

softmax is used for producing next-token probabilities

Supervised Finetuning of LMs

LMs are adapted to human preferences and downstream tasks via **finetuning**

Supervised Finetuning of LMs

LMs are adapted to human preferences and downstream tasks via **finetuning**

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs via **gradient-based methods**

Supervised Finetuning of LMs

LMs are adapted to human preferences and downstream tasks via **finetuning**

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs via **gradient-based methods**

outputs sampled from conditional distribution $\mathcal{D}(\cdot|\mathbf{x})$

Supervised Finetuning of LMs

LMs are adapted to human preferences and downstream tasks via **finetuning**

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs via **gradient-based methods**

outputs sampled from conditional distribution $\mathcal{D}(\cdot|\mathbf{x})$

Expected loss for input $\mathbf{x}$: $\mathcal{L}_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\mathcal{D}(\cdot|\mathbf{x})}\left[-\ln p_\theta(\mathbf{y}|\mathbf{x})\right]$

Supervised Finetuning of LMs

LMs are adapted to human preferences and downstream tasks via **finetuning**

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs via **gradient-based methods**

outputs sampled from conditional distribution $\mathcal{D}(\cdot|\mathbf{x})$

Expected loss for input $\mathbf{x}$: $\mathcal{L}_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\mathcal{D}(\cdot|\mathbf{x})}\left[-\ln p_\theta(\mathbf{y}|\mathbf{x})\right]$

Limitations:

Supervised Finetuning of LMs

LMs are adapted to human preferences and downstream tasks via **finetuning**

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs via **gradient-based methods**

outputs sampled from conditional distribution $\mathcal{D}(\cdot|\mathbf{x})$

Expected loss for input $\mathbf{x}$: $\mathcal{L}_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\mathcal{D}(\cdot|\mathbf{x})}\left[-\ln p_\theta(\mathbf{y}|\mathbf{x})\right]$

Limitations:

Hard to formalize human preferences through labels

Supervised Finetuning of LMs

LMs are adapted to human preferences and downstream tasks via **finetuning**

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs via **gradient-based methods**

outputs sampled from conditional distribution $\mathcal{D}(\cdot|\mathbf{x})$

Expected loss for input $\mathbf{x}$: $\mathcal{L}_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\mathcal{D}(\cdot|\mathbf{x})}\left[-\ln p_\theta(\mathbf{y}|\mathbf{x})\right]$

Limitations:

- Hard to formalize human preferences through labels
- Labeled data is expensive

Reinforcement Finetuning of LMs

Limitations of SFT led to wide adoption of a **reinforcement learning**-based approach

(e.g. Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022, Bai et al. 2022, Dubois et al. 2023, Touvron et al. 2023)

Reinforcement Finetuning of LMs

Limitations of SFT led to wide adoption of a **reinforcement learning**-based approach

(e.g. Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022, Bai et al. 2022, Dubois et al. 2023, Touvron et al. 2023)

Reinforcement Finetuning (RFT)

Maximize reward over unlabeled inputs via **policy gradient algorithms**

reward function $r(\mathbf{x}, \mathbf{y})$

Reinforcement Finetuning of LMs

Limitations of SFT led to wide adoption of a **reinforcement learning**-based approach

(e.g. Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022, Bai et al. 2022, Dubois et al. 2023, Touvron et al. 2023)

Reinforcement Finetuning (RFT)

Maximize reward over unlabeled inputs via **policy gradient algorithms**

reward function $r(\mathbf{x}, \mathbf{y})$

Expected reward for input $\mathbf{x}$: $V_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot|\mathbf{x})} [r(\mathbf{x}, \mathbf{y})]$

Reinforcement Finetuning of LMs

Limitations of SFT led to wide adoption of a **reinforcement learning**-based approach

(e.g. Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022, Bai et al. 2022, Dubois et al. 2023, Touvron et al. 2023)

Reinforcement Finetuning (RFT)

Maximize reward over unlabeled inputs via **policy gradient algorithms**

reward function $r(\mathbf{x}, \mathbf{y})$

Expected reward for input $\mathbf{x}$: $V_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot|\mathbf{x})} [r(\mathbf{x}, \mathbf{y})]$

Reward function $r(\mathbf{x}, \mathbf{y})$ can be:

Reinforcement Finetuning of LMs

Limitations of SFT led to wide adoption of a **reinforcement learning**-based approach

(e.g. Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022, Bai et al. 2022, Dubois et al. 2023, Touvron et al. 2023)

Reinforcement Finetuning (RFT)

Maximize reward over unlabeled inputs via **policy gradient algorithms**

, , $\cdots$, reward function $r(\mathbf{x}, \mathbf{y})$

Expected reward for input $\mathbf{x}$: $V_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot|\mathbf{x})} \left[r(\mathbf{x}, \mathbf{y}) \right]$

Reward function $r(\mathbf{x}, \mathbf{y})$ can be:

Learned from human preferences

Reinforcement Finetuning of LMs

Limitations of SFT led to wide adoption of a **reinforcement learning**-based approach

(e.g. Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022, Bai et al. 2022, Dubois et al. 2023, Touvron et al. 2023)

Reinforcement Finetuning (RFT)

Maximize reward over unlabeled inputs via **policy gradient algorithms**

reward function $r(\mathbf{x}, \mathbf{y})$

Expected reward for input $\mathbf{x}$: $V_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot|\mathbf{x})} [r(\mathbf{x}, \mathbf{y})]$

Reward function $r(\mathbf{x}, \mathbf{y})$ can be:

- Learned from human preferences
- Tailored to a downstream task

Main Contributions: Vanishing Gradients in RFT

Main Contributions: Vanishing Gradients in RFT

Fundamental vanishing gradients problem in RFT $\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Main Contributions: Vanishing Gradients in RFT

Fundamental vanishing gradients problem in RFT $\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Vanishing gradients are prevalent and harm ability to maximize reward

Main Contributions: Vanishing Gradients in RFT

Fundamental vanishing gradients problem in RFT $\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Vanishing gradients are prevalent and harm ability to maximize reward

Exploring ways to overcome vanishing gradients in RFT

Main Contributions: Vanishing Gradients in RFT

Fundamental vanishing gradients problem in RFT $\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Vanishing gradients are prevalent and harm ability to maximize reward

Exploring ways to overcome vanishing gradients in RFT

Vanishing Gradients Due to Small Reward Standard Deviation (STD)

$\mathrm{STD}_{\mathbf{y} \sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})]$ — reward std of $\mathbf{x}$ under the model

Vanishing Gradients Due to Small Reward Standard Deviation (STD)

$\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]$ — reward std of $\mathbf{x}$ under the model

Theorem

$$\|\nabla_\theta V_\theta(\mathbf{x})\| = O\big(\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]^{2/3}\big)$$

*Same holds for PPO gradient

Vanishing Gradients Due to Small Reward Standard Deviation (STD)

$\mathrm{STD}_{\mathbf{y} \sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})]$ — reward std of $\mathbf{x}$ under the model

Theorem

$$\|\nabla_\theta V_\theta(\mathbf{x})\| = O\left(\mathrm{STD}_{\mathbf{y} \sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})]^{2/3}\right)$$

Expected gradient for an input vanishes when reward std is small, even if reward mean is suboptimal

*Same holds for PPO gradient

Vanishing Gradients Due to Small Reward Standard Deviation (STD)

$\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]$ — reward std of $\mathbf{x}$ under the model

Theorem

$$\|\nabla_\theta V_\theta(\mathbf{x})\| = O\left(\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]^{2/3}\right)$$

Expected gradient for an input vanishes when reward std is small, even if reward mean is suboptimal

*Same holds for PPO gradient

Proof Idea: Stems from use of softmax + reward maximization objective

Vanishing Gradients Due to Small Reward Standard Deviation (STD)

$\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]$ — reward std of $\mathbf{x}$ under the model

Theorem

$$\|\nabla_\theta V_\theta(\mathbf{x})\| = O\left(\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]^{2/3}\right)$$

Expected gradient for an input vanishes when reward std is small, even if reward mean is suboptimal

*Same holds for PPO gradient

Proof Idea: Stems from use of softmax + reward maximization objective

Note: Bound applies to expected gradients of individual inputs (as opposed to of batch/population)

Vanishing Gradients Due to Small Reward Standard Deviation (STD)

$\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]$ — reward std of $\mathbf{x}$ under the model

Theorem

$$\|\nabla_\theta V_\theta(\mathbf{x})\| = O\left(\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]^{2/3}\right)$$

Expected gradient for an input vanishes when reward std is small, even if reward mean is suboptimal

*Same holds for PPO gradient

Proof Idea: Stems from use of softmax + reward maximization objective

Note: Bound applies to expected gradients of individual inputs (as opposed to of batch/population)

Can be problematic when finetuning text distribution differs from pretraining

Main Contributions: Vanishing Gradients in RFT

Fundamental vanishing gradients problem in RFT $\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Vanishing gradients are prevalent and harm ability to maximize reward

Exploring ways to overcome vanishing gradients in RFT

Main Contributions: Vanishing Gradients in RFT

Fundamental vanishing gradients problem in RFT $\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Vanishing gradients are prevalent and harm ability to maximize reward

Exploring ways to overcome vanishing gradients in RFT

Prevalence and Detrimental Effects of Vanishing Gradients

Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Finding I

3 of 7 datasets contain considerable # of train inputs with small reward std and low reward

Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Finding I

3 of 7 datasets contain considerable # of train inputs with $\overbrace{\text{small reward std}}^{\text{vanishing gradients}}$ and low reward

Prevalence and Detrimental Effects of Vanishing Gradients

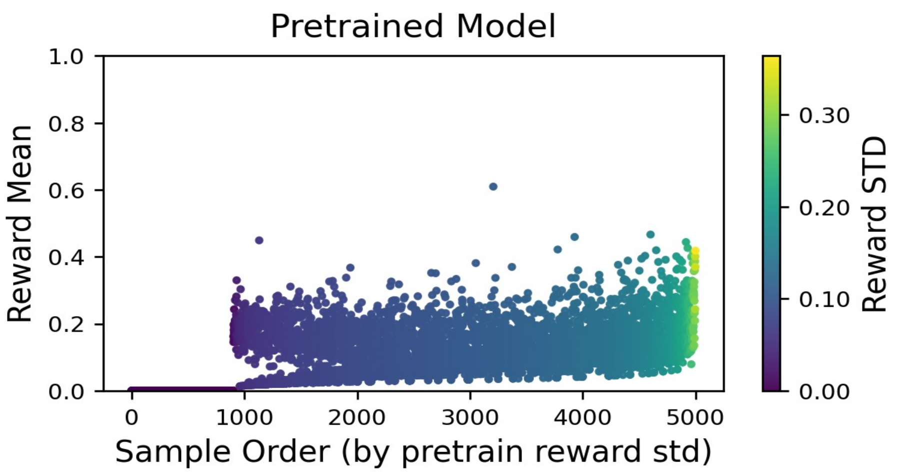
Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Finding I

3 of 7 datasets contain considerable # of train inputs with $\underbrace{\text{small reward std}}_{\text{vanishing gradients}}$ and low reward

NarrativeQA
(many inputs w/ small std)



Prevalence and Detrimental Effects of Vanishing Gradients

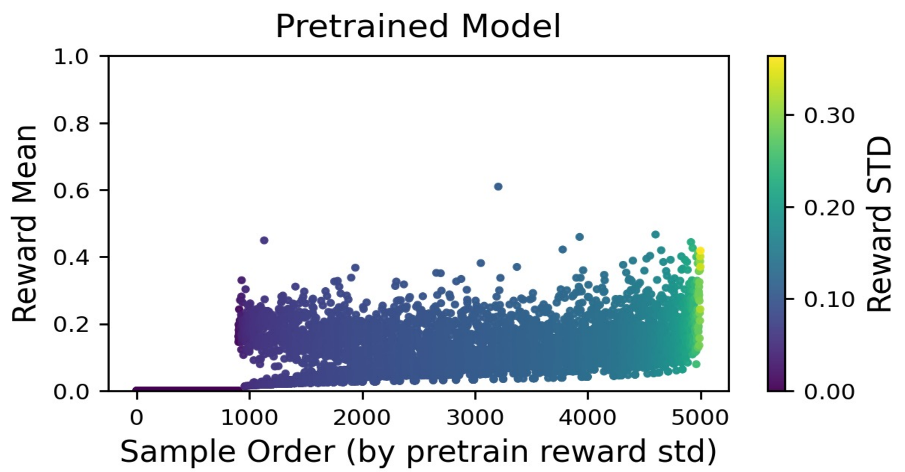
Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

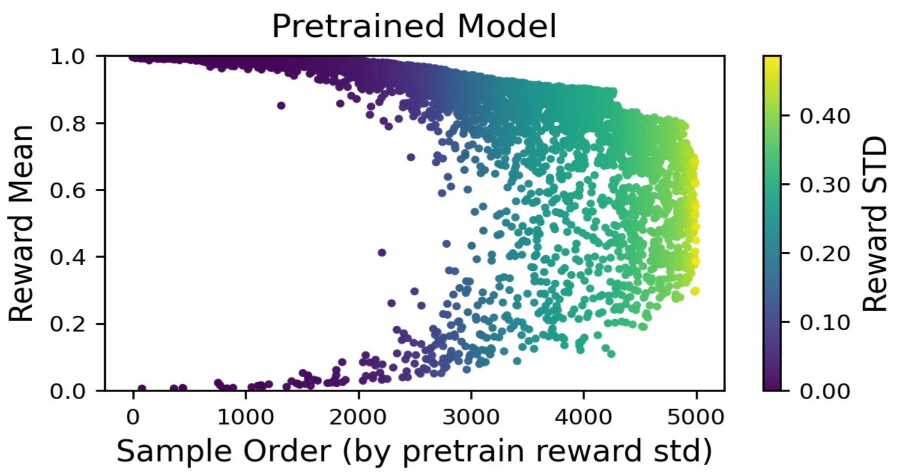
Finding I

3 of 7 datasets contain considerable # of train inputs with $\underbrace{\text{small reward std}}_{\text{vanishing gradients}}$ and low reward

NarrativeQA
(many inputs w/ small std)



IMDB
(few inputs w/ small std)



Which Datasets Suffer From Vanishing Gradients in RFT?

Which Datasets Suffer From Vanishing Gradients in RFT?

As expected: Text distribution substantially differs from pretraining distribution

Which Datasets Suffer From Vanishing Gradients in RFT?

As expected: Text distribution substantially differs from pretraining distribution

→ considerable amount of inputs with small reward std

Which Datasets Suffer From Vanishing Gradients in RFT?

As expected: Text distribution substantially differs from pretraining distribution

→ considerable amount of inputs with $\underbrace{\text{small reward std}}_{\text{vanishing gradients}}$

Which Datasets Suffer From Vanishing Gradients in RFT?

As expected: Text distribution substantially differs from pretraining distribution

→ considerable amount of inputs with $\underbrace{\text{small reward std}}_{\text{vanishing gradients}}$

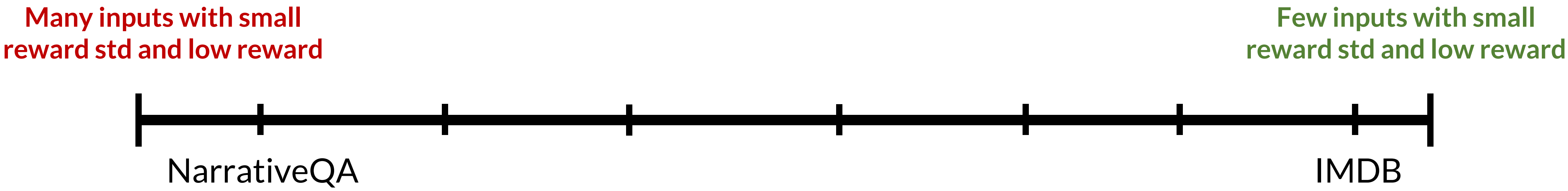
Many inputs with small reward std and low reward

Few inputs with small reward std and low reward

Which Datasets Suffer From Vanishing Gradients in RFT?

As expected: Text distribution substantially differs from pretraining distribution

→ considerable amount of inputs with $\underbrace{\text{small reward std}}_{\text{vanishing gradients}}$



Which Datasets Suffer From Vanishing Gradients in RFT?

As expected: Text distribution substantially differs from pretraining distribution

→ considerable amount of inputs with small reward std

vanishing gradients

Many inputs with small reward std and low reward

Few inputs with small reward std and low reward

NarrativeQA ToTTo CommonGen IWSLT 2017 CNN DailyDialog IMDB

Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Finding II

As expected, RFT has limited impact on the reward of inputs with small reward std

Prevalence and Detrimental Effects of Vanishing Gradients

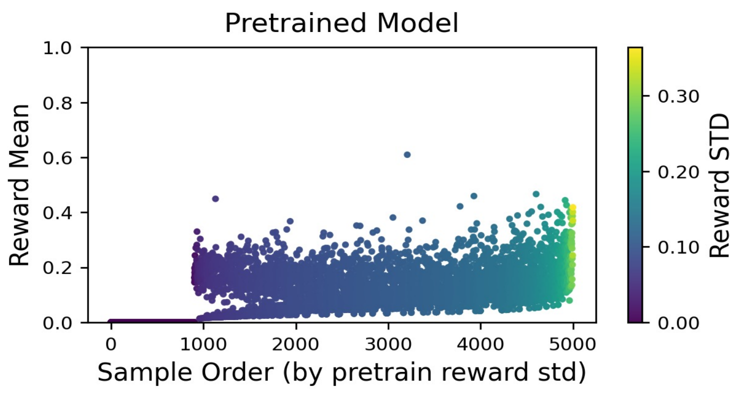
Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

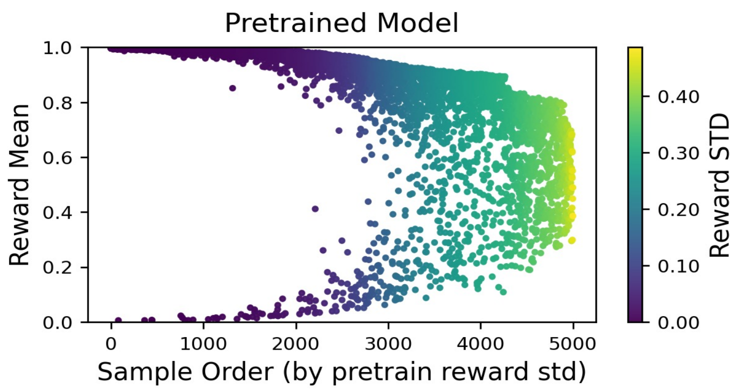
Finding II

As expected, RFT has limited impact on the reward of inputs with small reward std

NarrativeQA
(many inputs w/ small std)



IMDB
(few inputs w/ small std)



Prevalence and Detrimental Effects of Vanishing Gradients

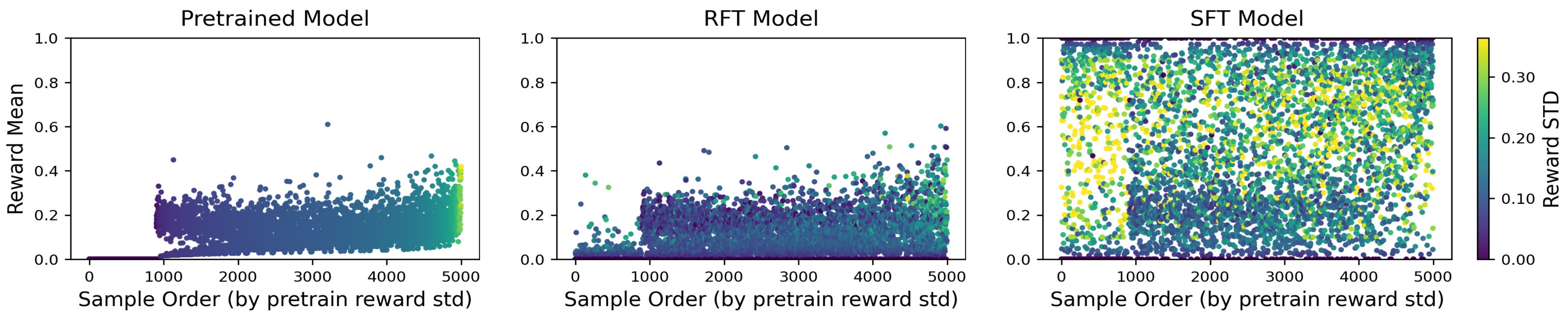
Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

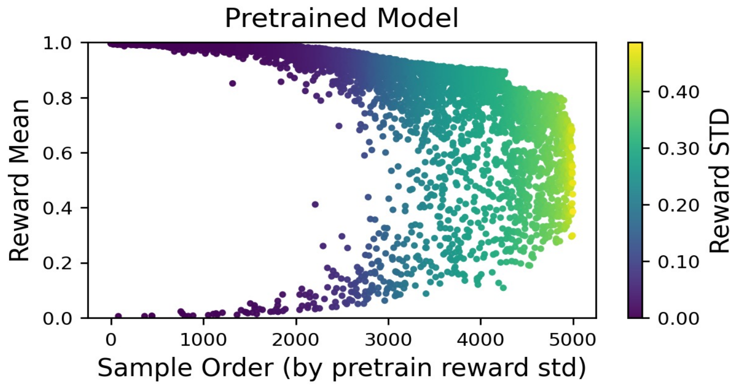
Finding II

As expected, RFT has limited impact on the reward of inputs with small reward std

NarrativeQA
(many inputs w/ small std)



IMDB
(few inputs w/ small std)



Prevalence and Detrimental Effects of Vanishing Gradients

<u>Benchmark</u>: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

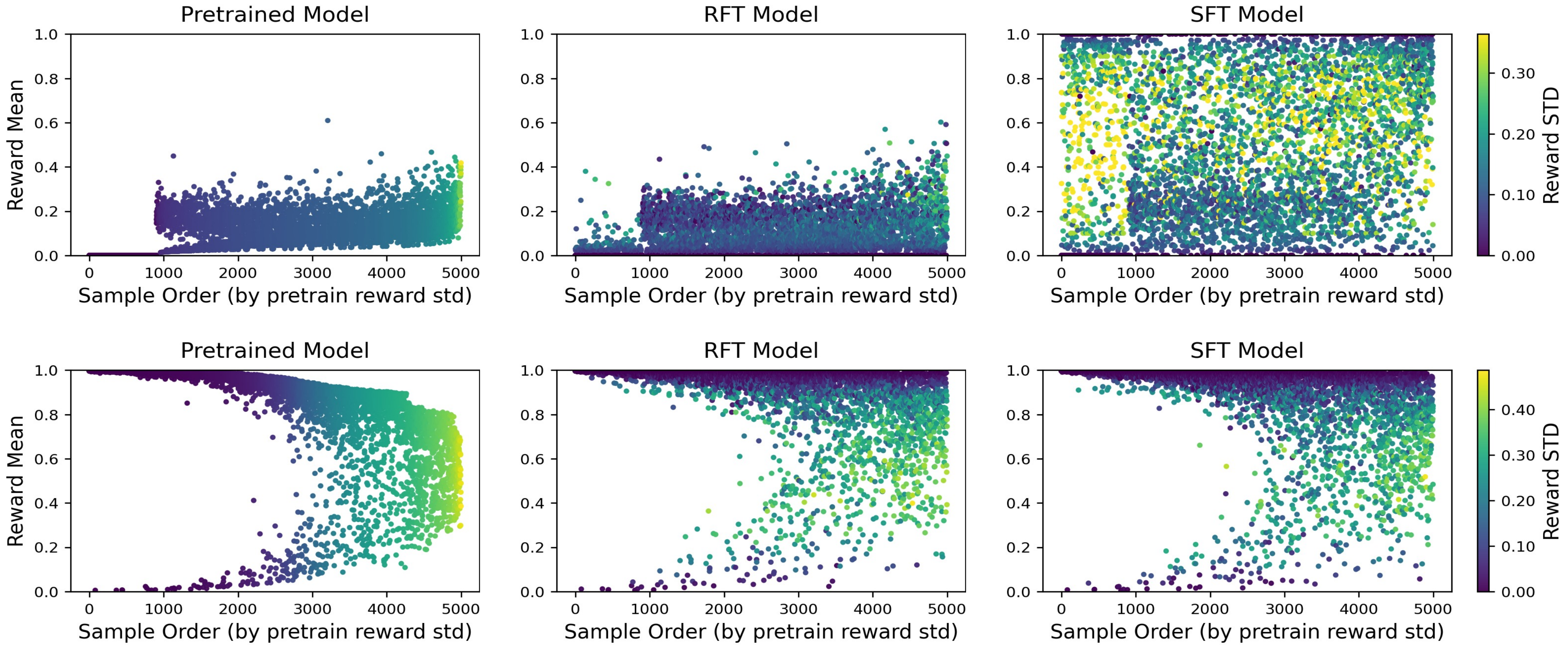
<u>Models</u>: GPT-2 and T5-base

Finding II

As expected, RFT has limited impact on the reward of inputs with small reward std

NarrativeQA
(many inputs w/ small std)

IMDB
(few inputs w/ small std)



Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Finding III

RFT performance is worse when inputs with small reward std are prevalent

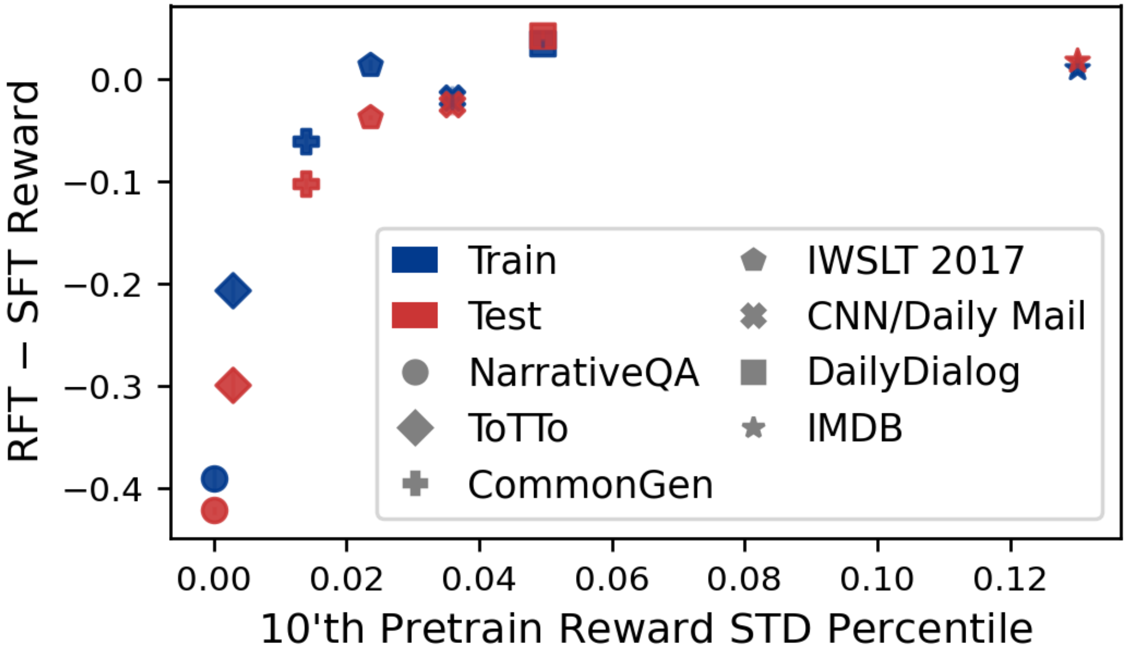
Prevalence and Detrimental Effects of Vanishing Gradients

Benchmark: GRUE (Ramamurthy et al. 2023)
7 language generation datasets

Models: GPT-2 and T5-base

Finding III

RFT performance is worse when inputs with small reward std are prevalent



Vanishing Gradients or Insufficient Exploration?

We saw that **vanishing expected gradients is indicative of RFT performance**

Vanishing Gradients or Insufficient Exploration?

We saw that $\underbrace{\textbf{vanishing expected gradients}}_{\text{measured by reward std}}$ **is indicative of RFT performance**

Vanishing Gradients or Insufficient Exploration?

We saw that $\underbrace{\textbf{vanishing expected gradients}}_{\text{measured by reward std}}$ **is indicative of RFT performance**

Possible Confounding Factor: Insufficient Exploration

Vanishing Gradients or Insufficient Exploration?

We saw that $\underbrace{\textbf{vanishing expected gradients}}_{\text{measured by reward std}}$ **is indicative of RFT performance**

Possible Confounding Factor: Insufficient Exploration

Large output space in language generation

Vanishing Gradients or Insufficient Exploration?

We saw that **vanishing expected gradients is indicative of RFT performance**

measured by reward std

Possible Confounding Factor: Insufficient Exploration

Large output space in language generation → challenge of exploration
(e.g. Ranzato et al. 2016, Choshen et al. 2020)

Vanishing Gradients or Insufficient Exploration?

We saw that $\underbrace{\textbf{vanishing expected gradients}}_{\text{measured by reward std}}$ **is indicative of RFT performance**

Possible Confounding Factor: Insufficient Exploration

Large output space in language generation $\longrightarrow$ challenge of exploration
(e.g. Ranzato et al. 2016, Choshen et al. 2020)

$\longrightarrow$ challenge of accurately estimating $\nabla_\theta V_\theta(\mathbf{x})$

Vanishing Gradients or Insufficient Exploration?

We saw that $\underbrace{\textbf{vanishing expected gradients}}_{\text{measured by reward std}}$ **is indicative of RFT performance**

Possible Confounding Factor: Insufficient Exploration

Large output space in language generation → challenge of exploration
(e.g. Ranzato et al. 2016, Choshen et al. 2020)

→ challenge of accurately estimating $\nabla_\theta V_\theta(\mathbf{x})$

Q: Does the difficulty of RFT to maximize reward stem from vanishing gradients or just insufficient exploration?

Vanishing Gradients or Insufficient Exploration?

We saw that $\underbrace{\textbf{vanishing expected gradients}}_{\text{measured by reward std}}$ **is indicative of RFT performance**

Possible Confounding Factor: Insufficient Exploration

Large output space in language generation → challenge of exploration
(e.g. Ranzato et al. 2016, Choshen et al. 2020)

→ challenge of accurately estimating $\nabla_\theta V_\theta(\mathbf{x})$

Q: Does the difficulty of RFT to maximize reward stem from vanishing gradients or just insufficient exploration?

ⓘ **We address Q via controlled experiments and theoretical analysis**

Controlled Experiments and Theoretical Analysis

Controlled Experiments and Theoretical Analysis

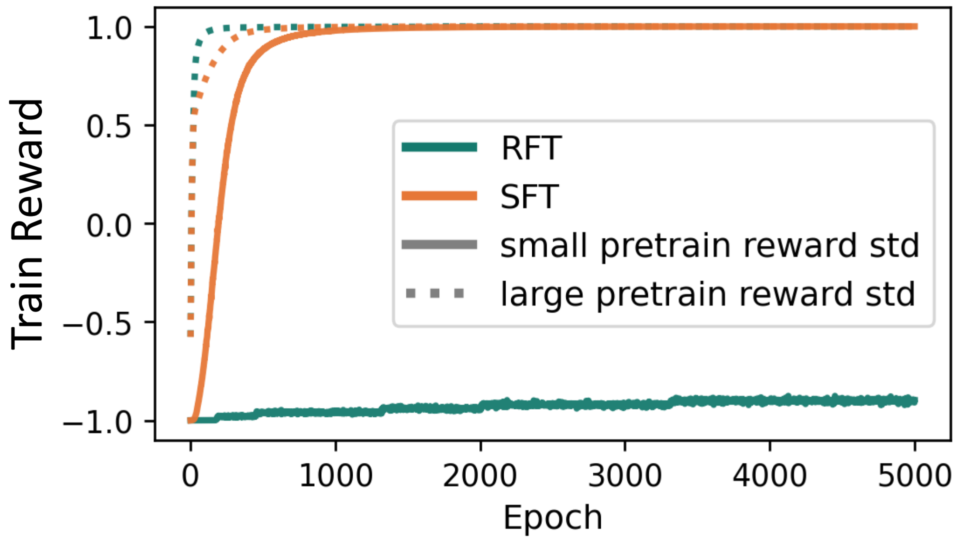
Controlled Experiments

Environments with **perfect exploration**,
i.e. RFT has access to expected gradients

Controlled Experiments and Theoretical Analysis

Controlled Experiments

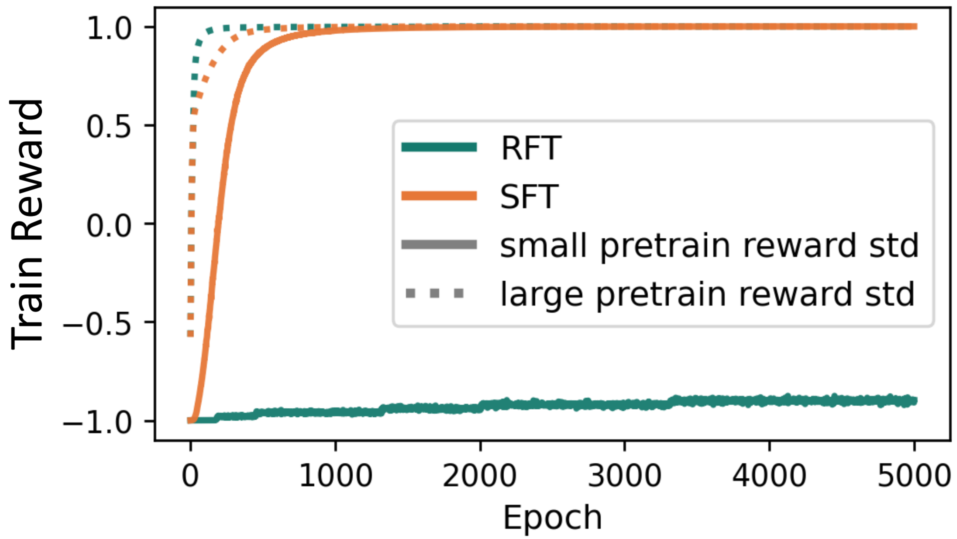
Environments with **perfect exploration**,
i.e. RFT has access to expected gradients



Controlled Experiments and Theoretical Analysis

Controlled Experiments

Environments with **perfect exploration**, i.e. RFT has access to expected gradients



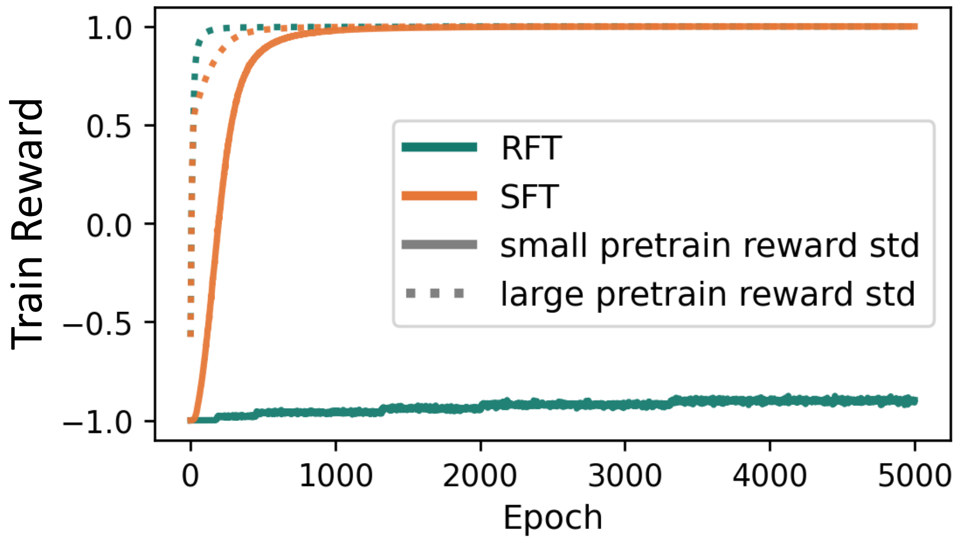
Theoretical Analysis

Simplified setting of linear classification over orthonormal data with **perfect exploration**

Controlled Experiments and Theoretical Analysis

Controlled Experiments

Environments with **perfect exploration**, i.e. RFT has access to expected gradients



Theoretical Analysis

Simplified setting of linear classification over orthonormal data with **perfect exploration**

Theorem

Time it takes to correctly classify input $\mathbf{x}$ is:

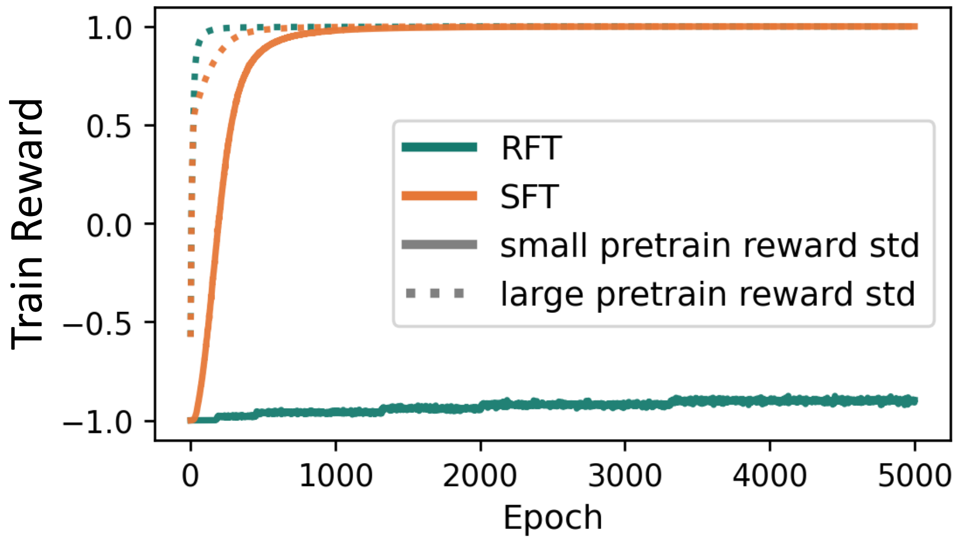
in RFT - $\Omega\left(1/\mathrm{STD}_{\mathbf{y}\sim p_{\theta(0)}(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]^2\right)$

in SFT - $O\left(\ln\left(1/\mathrm{STD}_{\mathbf{y}\sim p_{\theta(0)}(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]\right)\right)$

Controlled Experiments and Theoretical Analysis

Controlled Experiments

Environments with **perfect exploration**, i.e. RFT has access to expected gradients



Theoretical Analysis

Simplified setting of linear classification over orthonormal data with **perfect exploration**

Theorem

Time it takes to correctly classify input $\mathbf{x}$ is:

in RFT - $\Omega\left(1/\mathrm{STD}_{\mathbf{y}\sim p_{\theta(0)}(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]^2\right)$

in SFT - $O\left(\ln\left(1/\mathrm{STD}_{\mathbf{y}\sim p_{\theta(0)}(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]\right)\right)$

RFT struggles to maximize reward for inputs with small reward std despite perfect exploration

Main Contributions: Vanishing Gradients in RFT

Fundamental vanishing gradients problem in RFT $\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Vanishing gradients are prevalent and harm ability to maximize reward

Exploring ways to overcome vanishing gradients in RFT

Main Contributions: Vanishing Gradients in RFT

Fundamental vanishing gradients problem in RFT

$$\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$$

Vanishing gradients are prevalent and harm ability to maximize reward

Exploring ways to overcome vanishing gradients in RFT

Inadequacy of Common Heuristics

Vanishing gradients in RFT are resilient to common heuristics:

Inadequacy of Common Heuristics

Vanishing gradients in RFT are resilient to common heuristics:

- Increasing learning rate

Inadequacy of Common Heuristics

Vanishing gradients in RFT are resilient to common heuristics:

- Increasing learning rate
- Adding temperature to logits

Inadequacy of Common Heuristics

Vanishing gradients in RFT are resilient to common heuristics:

- Increasing learning rate
- Adding temperature to logits
- Entropy regularization

Inadequacy of Common Heuristics

Vanishing gradients in RFT are resilient to common heuristics:

- Increasing learning rate
- Adding temperature to logits
- Entropy regularization

Expected to help?

Inadequacy of Common Heuristics

Vanishing gradients in RFT are resilient to common heuristics:

- Increasing learning rate
- Adding temperature to logits
- Entropy regularization

Expected to help? ❌

Inadequacy of Common Heuristics

Vanishing gradients in RFT are resilient to common heuristics:

- Increasing learning rate
- Adding temperature to logits
- Entropy regularization

Expected to help? ✗

Results: As expected, no improvement to the reward of RFT

Inadequacy of Common Heuristics

Dataset: NarrativeQA

	Train Reward	Test Reward
RFT*	0.101 ± 0.009	0.116 ± 0.000
SFT + RFT	$\mathbf{0.537} \pm \mathbf{0.005}$	$\mathbf{0.544} \pm \mathbf{0.003}$
RFT with learning rate $2 \cdot 10^{-5}$	0.012 ± 0.010	0.020 ± 0.017
RFT with learning rate $2 \cdot 10^{-4}$	0.053 ± 0.010	0.048 ± 0.016
RFT with learning rate $2 \cdot 10^{-3}$	0.039 ± 0.018	0.012 ± 0.020
RFT with temperature 1.5	0.077 ± 0.021	0.118 ± 0.002
RFT with temperature 2	0.060 ± 0.018	0.104 ± 0.009
RFT with temperature 2.5	0.044 ± 0.004	0.088 ± 0.009
RFT with entropy regularization 0.01	0.080 ± 0.006	0.113 ± 0.002
RFT with entropy regularization 0.1	0.024 ± 0.005	0.019 ± 0.007
RFT with entropy regularization 1	0.011 ± 0.000	0.013 ± 0.010

*With default hyperparameters: learning rate $2 \cdot 10^{-6}$, temperature 1, entropy regularization 0

Initial SFT Phase Mitigates Vanishing Gradients in RFT

Common practice is to perform initial SFT phase before RFT (e.g. Ouyang et al. 2022)

Initial SFT Phase Mitigates Vanishing Gradients in RFT

Common practice is to perform initial SFT phase before RFT (e.g. Ouyang et al. 2022)

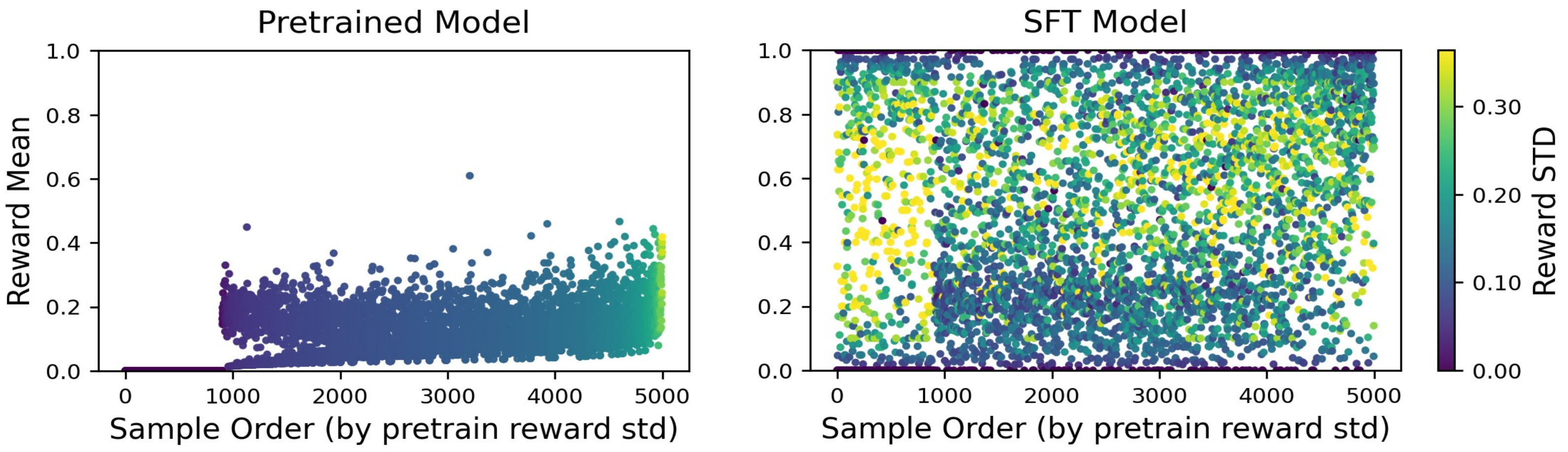
Observation – Initial SFT phase reduces number of inputs with small reward std

Initial SFT Phase Mitigates Vanishing Gradients in RFT

Common practice is to perform initial SFT phase before RFT (e.g. Ouyang et al. 2022)

Observation – Initial SFT phase reduces number of inputs with small reward std

NarrativeQA (train)

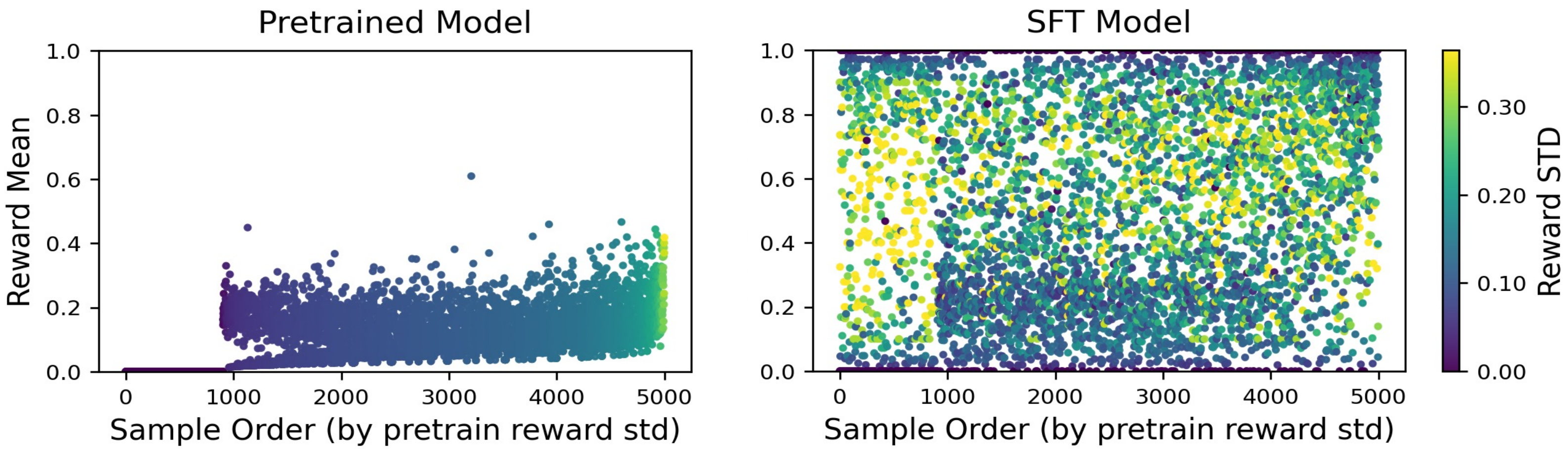


Initial SFT Phase Mitigates Vanishing Gradients in RFT

Common practice is to perform initial SFT phase before RFT (e.g. Ouyang et al. 2022)

Observation – Initial SFT phase reduces number of inputs with small reward std

NarrativeQA (train)



Importance of SFT in RFT pipeline: mitigates vanishing gradients

A Few SFT Steps on a Small Number of Samples Suffice

A Few SFT Steps on a Small Number of Samples Suffice

Limitation of initial SFT phase – Requires labeled data

A Few SFT Steps on a Small Number of Samples Suffice

Limitation of initial SFT phase – Requires labeled data

Expectation – If SFT phase is beneficial due to mitigating vanishing gradients for RFT

A Few SFT Steps on a Small Number of Samples Suffice

Limitation of initial SFT phase – Requires labeled data

Expectation – If SFT phase is beneficial due to mitigating vanishing gradients for RFT

→ A few steps of SFT on small # of labeled samples should suffice

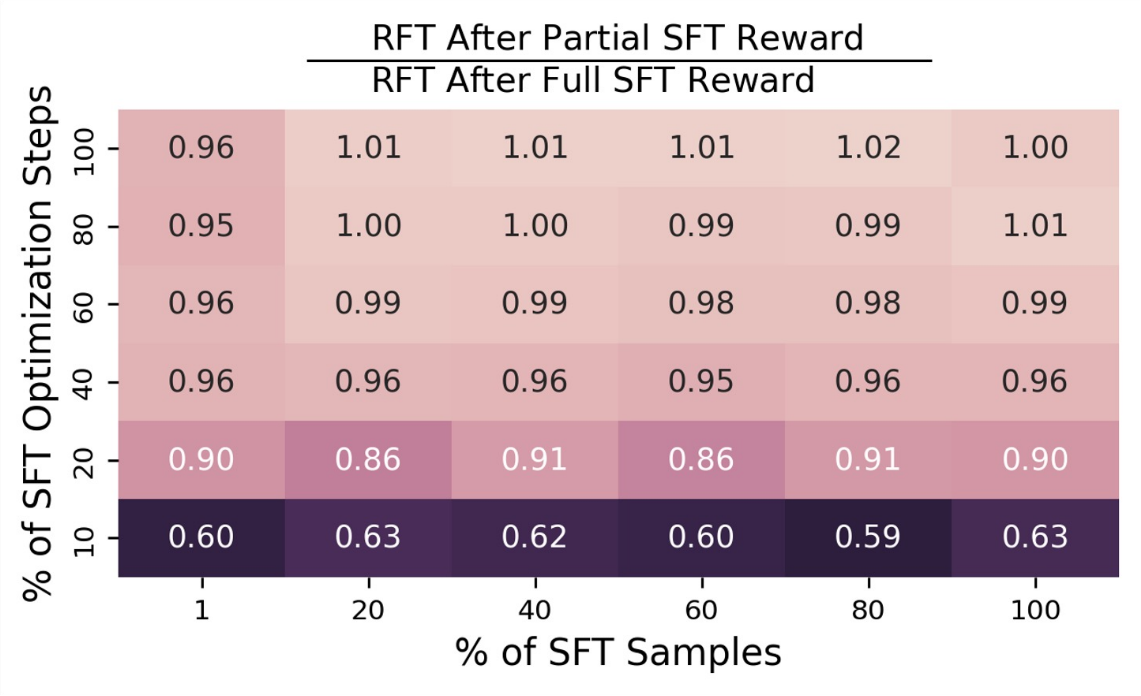
A Few SFT Steps on a Small Number of Samples Suffice

Limitation of initial SFT phase – Requires labeled data

Expectation – If SFT phase is beneficial due to mitigating vanishing gradients for RFT

→ A few steps of SFT on small # of labeled samples should suffice

NarrativeQA (train)

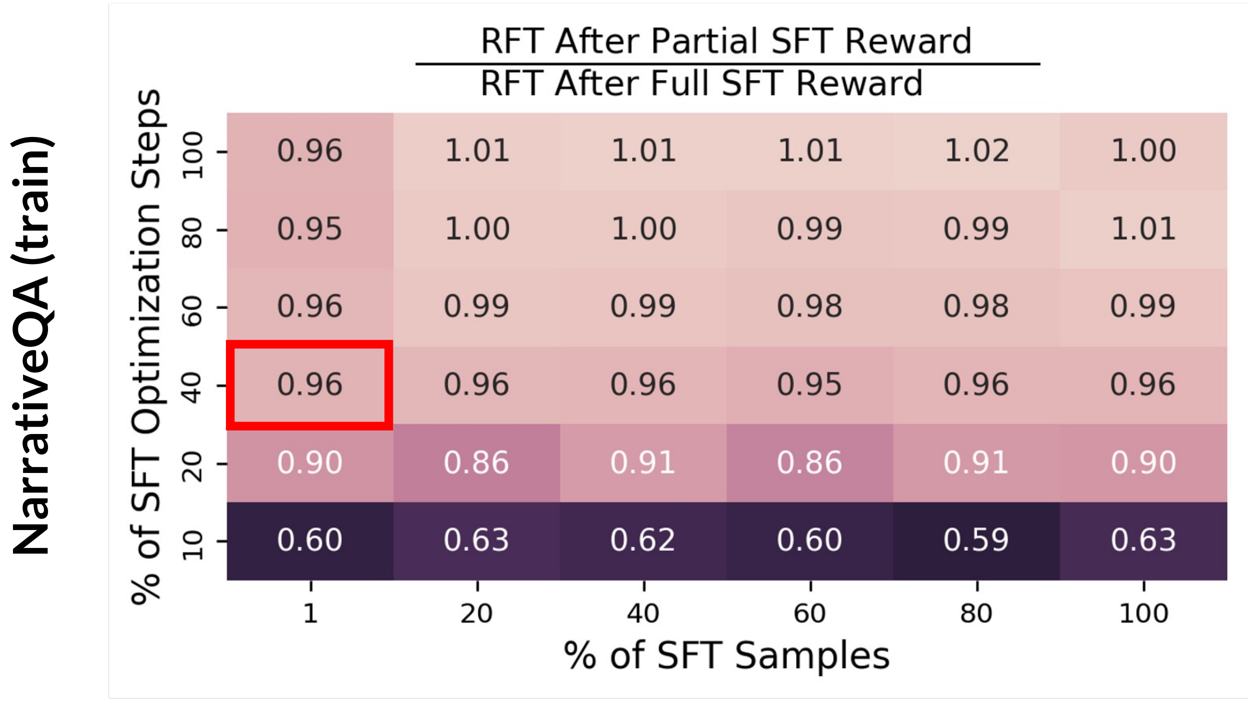


A Few SFT Steps on a Small Number of Samples Suffice

Limitation of initial SFT phase – Requires labeled data

Expectation – If SFT phase is beneficial due to mitigating vanishing gradients for RFT

→ A few steps of SFT on small # of labeled samples should suffice

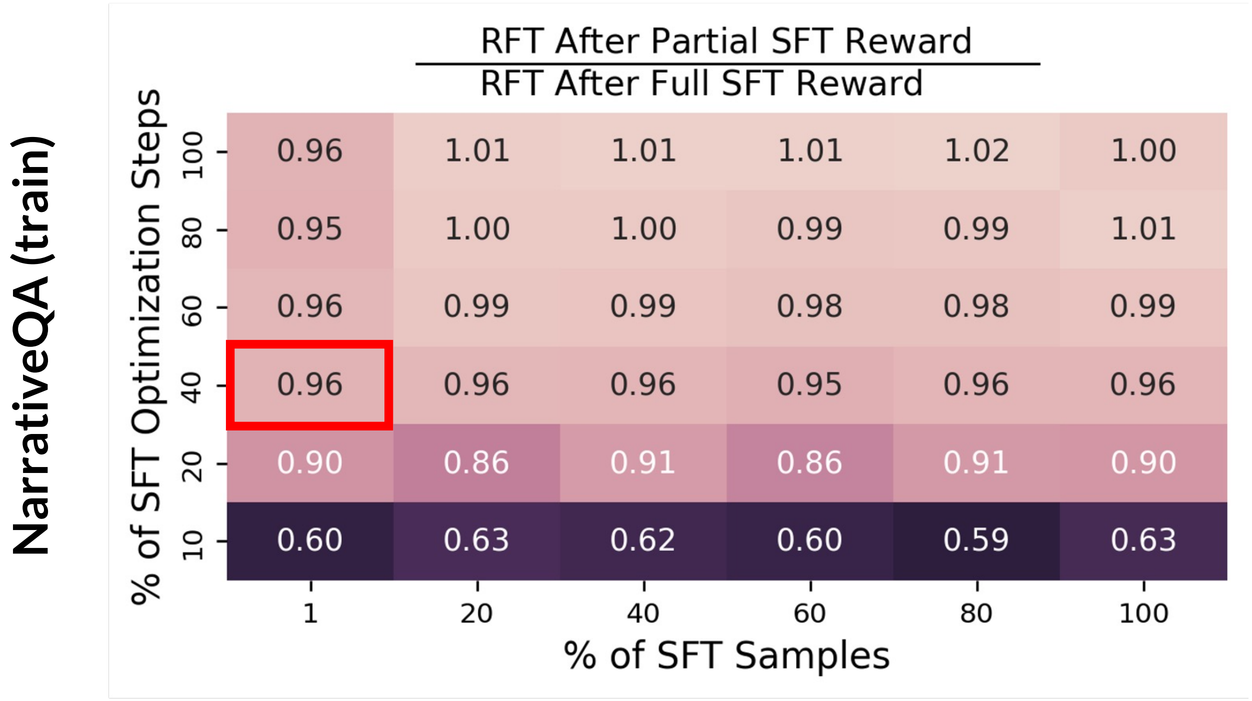


A Few SFT Steps on a Small Number of Samples Suffice

Limitation of initial SFT phase – Requires labeled data

Expectation – If SFT phase is beneficial due to mitigating vanishing gradients for RFT

→ A few steps of SFT on small # of labeled samples should suffice



The initial SFT phase does not need to be expensive!

Conclusion

Conclusion

$$\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$$

Expected gradient for an input vanishes in RFT
if the input's reward std is small

Conclusion

$\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Expected gradient for an input vanishes in RFT
if the input's reward std is small

Experiments + theory: **vanishing gradients in RFT are prevalent and detrimental** to maximizing reward

Conclusion

$\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$ **Expected gradient for an input vanishes in RFT** if the input's reward std is small

Experiments + theory: **vanishing gradients in RFT are prevalent and detrimental** to maximizing reward

Initial SFT phase allows overcoming vanishing gradients in RFT, and **does not need to be expensive**

Conclusion

$\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$ **Expected gradient for an input vanishes in RFT** if the input's reward std is small

Experiments + theory: **vanishing gradients in RFT are prevalent and detrimental** to maximizing reward

Initial SFT phase allows overcoming vanishing gradients in RFT, and **does not need to be expensive**

Reward std is a key quantity to track for successful RFT

Conclusion

Thank You!

$\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$ **Expected gradient for an input vanishes in RFT** if the input's reward std is small

Experiments + theory: **vanishing gradients in RFT are prevalent and detrimental** to maximizing reward

Initial SFT phase allows overcoming vanishing gradients in RFT, and **does not need to be expensive**

Reward std is a key quantity to track for successful RFT