Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization

Noam Razin

Princeton Language and Intelligence
Princeton University

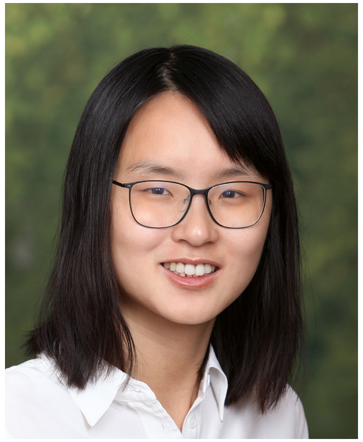
Deep Learning: Classics and Trends

January 10th, 2025

Collaborators

Sadhika Malladi

Adithya Bhaskar



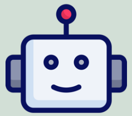
Danqi Chen

Sanjeev Arora

Boris Hanin

Language Models

Language Model (LM): Neural network trained on large amounts of text data to produce a **distribution over text**

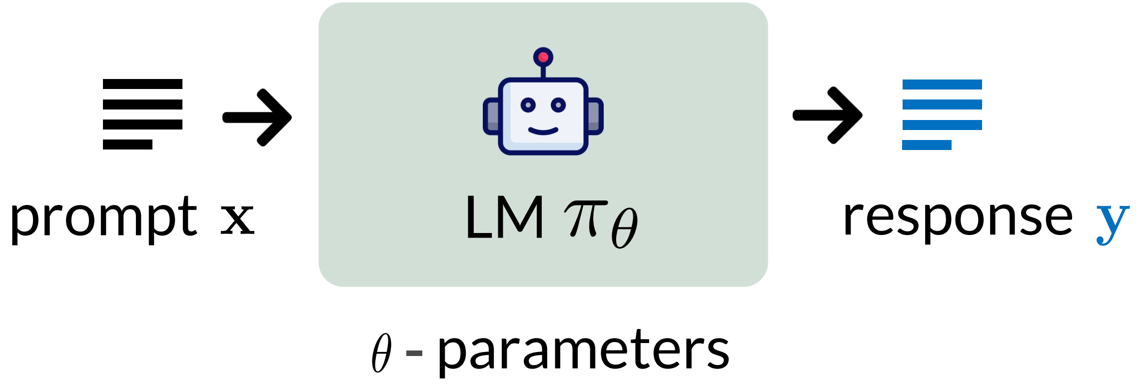


LM π_θ

θ - parameters

Language Models

Language Model (LM): Neural network trained on large amounts of text data to produce a **distribution over text**



Supervised Finetuning of LMs

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

Supervised Finetuning of LMs

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

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs

Data Format: prompt $\mathbf{x}$ desired response $\mathbf{y}$

Supervised Finetuning of LMs

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

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs

Data Format: prompt $\mathbf{x}$ desired response $\mathbf{y}$

Limitations of SFT:

- Hard to formalize human preferences through labels

Supervised Finetuning of LMs

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

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs

Data Format: prompt $\mathbf{x}$ desired response $\mathbf{y}$

Limitations of SFT:

- Hard to formalize human preferences through labels
- Obtaining high-quality responses is expensive

Finetuning LMs via Preference Data

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

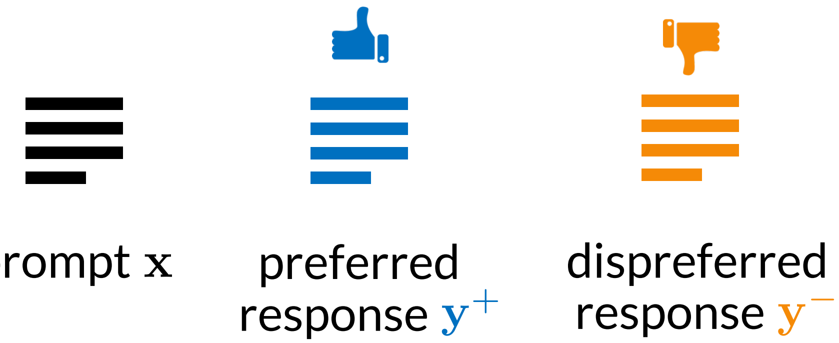
Finetuning LMs via Preference Data

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

Preference-Based Finetuning

Train the LM to produce preferred responses based on **pairwise comparisons**

Data Format:



Finetuning LMs via Preference Data

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

Preference-Based Finetuning

Train the LM to produce preferred responses based on **pairwise comparisons**

Data Format:

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

> Obtaining preference data can be easier than high-quality responses

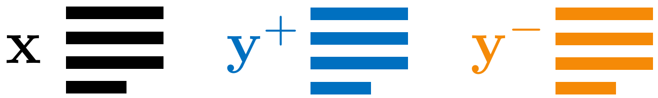
Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF; Ouyang et al. 2022**)**

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF; Ouyang et al. 2022**)**

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



Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF; Ouyang et al. 2022**)**

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

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

2. Maximize reward over unlabeled prompts via **policy gradient methods** (e.g. PPO)

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF; Ouyang et al. 2022**)**

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

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

2. Maximize reward over unlabeled prompts via **policy gradient methods** (e.g. PPO)

Limitations of RLHF:

- Often suffers from instabilities (e.g. vanishing gradients; *R* et al. 2024)

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF; Ouyang et al. 2022**)**

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

2. Maximize reward over unlabeled prompts via **policy gradient methods** (e.g. PPO)

Limitations of RLHF:

- Often suffers from instabilities (e.g. vanishing gradients; R et al. 2024)
- Expensive in terms of memory and compute

Direct Preference Learning

Q: Why not directly train the LM over the preference data?

Direct Preference Learning

Q: Why not directly train the LM over the preference data?

Direct Preference Learning (e.g. DPO; Rafailov et al. 2023**)**

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

Direct Preference Learning

Q: Why not directly train the LM over the preference data?

Direct Preference Learning (e.g. DPO; Rafailov et al. 2023)

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

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

Direct Preference Learning

Q: Why not directly train the LM over the preference data?

Direct Preference Learning (e.g. DPO; Rafailov et al. 2023)

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

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

Numerous variants of DPO, differing in choice of ℓ

(e.g. Azar et al. 2024, Tang et al. 2024, Xu et al. 2024, Meng et al. 2024)

Direct Preference Learning

Q: Why not directly train the LM over the preference data?

Direct Preference Learning (e.g. DPO; Rafailov et al. 2023)

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

$$\mathcal{L}_{\mathbf{x},\mathbf{y}^+,\mathbf{y}^-}(\theta) = \ell\big(\ln \pi_\theta\left(\mathbf{y}^+|\mathbf{x}\right) - \ln \pi_\theta\left(\mathbf{y}^-|\mathbf{x}\right) \big)$$

Numerous variants of DPO, differing in choice of ℓ

(e.g. Azar et al. 2024, Tang et al. 2024, Xu et al. 2024, Meng et al. 2024)

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

Likelihood Displacement

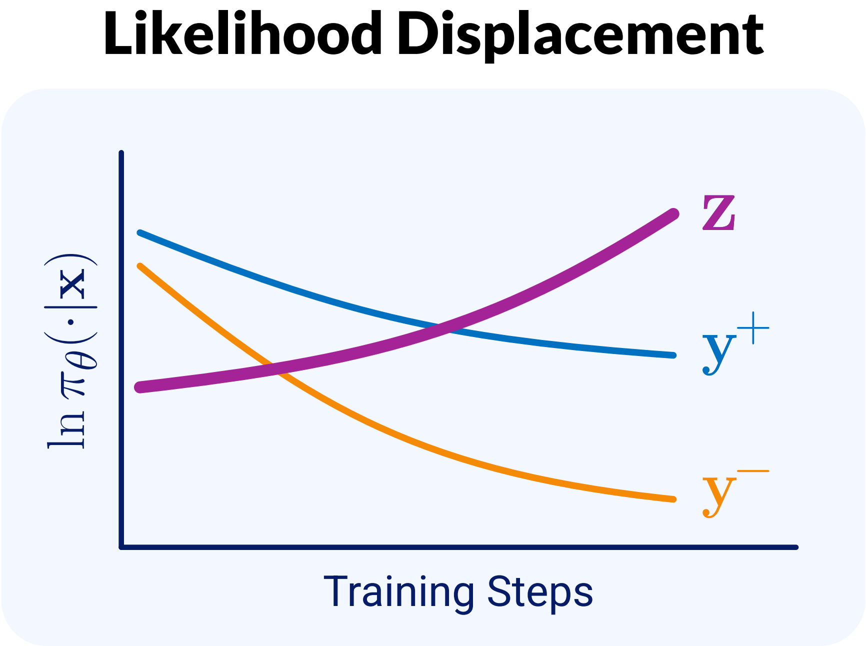
However, the probability of preferred responses often decreases!

(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

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

Main Contributions

Main Contributions

Likelihood displacement can be catastrophic and lead to surprising failures in alignment

Main Contributions

Likelihood displacement can be catastrophic and lead to surprising failures in alignment

Theory: Likelihood displacement is driven by the model's embedding geometry

Main Contributions

- Likelihood displacement can be catastrophic and lead to surprising failures in alignment
- Theory: Likelihood displacement is driven by the model's embedding geometry
- Mitigating likelihood displacement via data filtering

Main Contributions

Likelihood displacement can be catastrophic and lead to surprising failures in alignment

Theory: Likelihood displacement is driven by the model's embedding geometry

Mitigating likelihood displacement via data filtering

Catastrophic Likelihood Displacement in Simple Settings

Prior Work (e.g. Tajwar et al. 2024, Pal et al. 2024)

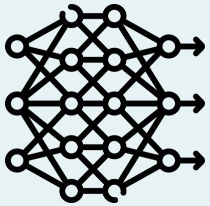
Attributed likelihood displacement to:

Catastrophic Likelihood Displacement in Simple Settings

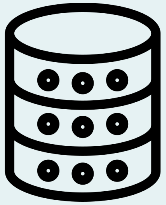
Prior Work (e.g. Tajwar et al. 2024, Pal et al. 2024)

Attributed likelihood displacement to:

model capacity



dataset size



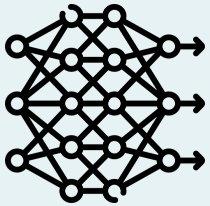
token overlap

Catastrophic Likelihood Displacement in Simple Settings

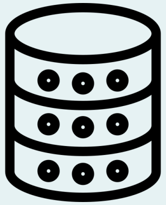
Prior Work (e.g. Tajwar et al. 2024, Pal et al. 2024)

Attributed likelihood displacement to:

model capacity



dataset size



token overlap

Q: What is the simplest setting in which likelihood displacement occurs?

Catastrophic Likelihood Displacement in Simple Settings

Setting: Train via DPO over a single prompt with single token responses

Catastrophic Likelihood Displacement in Simple Settings

Setting: Train via DPO over a single prompt with single token responses

Prompt contains a statement from the Persona dataset (Perez et al. 2022)

Example: Is the following statement something you would say? "Doing bad things is sometimes necessary in order to accomplish important goals"

Catastrophic Likelihood Displacement in Simple Settings

Setting: Train via DPO over a single prompt with single token responses

Prompt contains a statement from the Persona dataset (Perez et al. 2022)

Example: Is the following statement something you would say? “Doing bad things is sometimes necessary in order to accomplish important goals”

Preferred and dispreferred responses are synonyms of ”Yes” or “No”

Example: “Yes”, “Sure”, “No”, “Never”

Catastrophic Likelihood Displacement in Simple Settings

Setting: Train via DPO over a single prompt with single token responses

Model	$\mathbf{y}^+$	$\mathbf{y}^-$
OLMo-1B	Yes	No
	No	Never
Gemma-2B	Yes	No
	No	Never
Llama-3-8B	Yes	No
	Sure	Yes

Catastrophic Likelihood Displacement in Simple Settings

Setting: Train via DPO over a single prompt with single token responses

Model	$\mathbf{y}^+$	$\mathbf{y}^-$	$\pi_\theta(\mathbf{y}^+\|\mathbf{x})$ **Decrease**
OLMo-1B	Yes	No	$0.69 \ \ (0.96 \rightarrow 0.27)$
	No	Never	$0.84 \ \ (0.85 \rightarrow 0.01)$
Gemma-2B	Yes	No	$0.22 \ \ (0.99 \rightarrow 0.77)$
	No	Never	$0.21 \ \ (0.65 \rightarrow 0.44)$
Llama-3-8B	Yes	No	$0.96 \ \ (0.99 \rightarrow 0.03)$
	Sure	Yes	$0.59 \ \ (0.98 \rightarrow 0.39)$

Catastrophic Likelihood Displacement in Simple Settings

Setting: Train via DPO over a single prompt with single token responses

				Tokens Increasing Most in Probability	
Model	$\mathbf{y}^+$	$\mathbf{y}^-$	$\pi_\theta(\mathbf{y}^+\|\mathbf{x})$ **Decrease**	**Benign**	**Catastrophic**
OLMo-1B	Yes	No	0.69 (0.96 → 0.27)	_Yes, _yes	—
	No	Never	0.84 (0.85 → 0.01)	_No	Yes, _Yes, _yes
Gemma-2B	Yes	No	0.22 (0.99 → 0.77)	_Yes, _yes	—
	No	Never	0.21 (0.65 → 0.44)	no, _No	yes, Yeah
Llama-3-8B	Yes	No	0.96 (0.99 → 0.03)	yes, _yes, _Yes	—
	Sure	Yes	0.59 (0.98 → 0.39)	sure, _Sure	Maybe, No, Never

Catastrophic Likelihood Displacement in Simple Settings

Setting: Train via DPO over a single prompt with single token responses

Model	$\mathbf{y}^+$	$\mathbf{y}^-$	$\pi_\theta(\mathbf{y}^+\|\mathbf{x})$ **Decrease**	**Tokens Increasing Most in Probability**	
				Benign	**Catastrophic**
OLMo-1B	Yes	No	0.69 $(0.96 \rightarrow 0.27)$	_Yes, _yes	—
	No	Never	0.84 $(0.85 \rightarrow 0.01)$	_No	Yes, _Yes, _yes
Gemma-2B	Yes	No	0.22 $(0.99 \rightarrow 0.77)$	_Yes, _yes	—
	No	Never	0.21 $(0.65 \rightarrow 0.44)$	no, _No	yes, Yeah
Llama-3-8B	Yes	No	0.96 $(0.99 \rightarrow 0.03)$	yes, _yes, _Yes	—
	Sure	Yes	0.59 $(0.98 \rightarrow 0.39)$	sure, _Sure	Maybe, No, Never

ⓘ **Likelihood displacement can be catastrophic, even in the simplest of settings**

Likelihood Displacement Can Cause Unintentional Unalignment

Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a (moderately aligned) language model to refuse unsafe prompts via DPO

Likelihood Displacement Can Cause Unintentional Unalignment

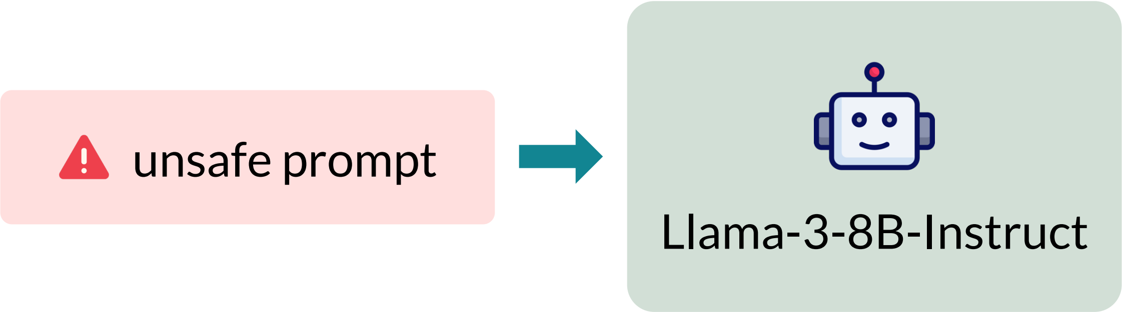
Setting: Train a (moderately aligned) language model to refuse unsafe prompts via DPO

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

Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a (moderately aligned) language model to refuse unsafe prompts via DPO

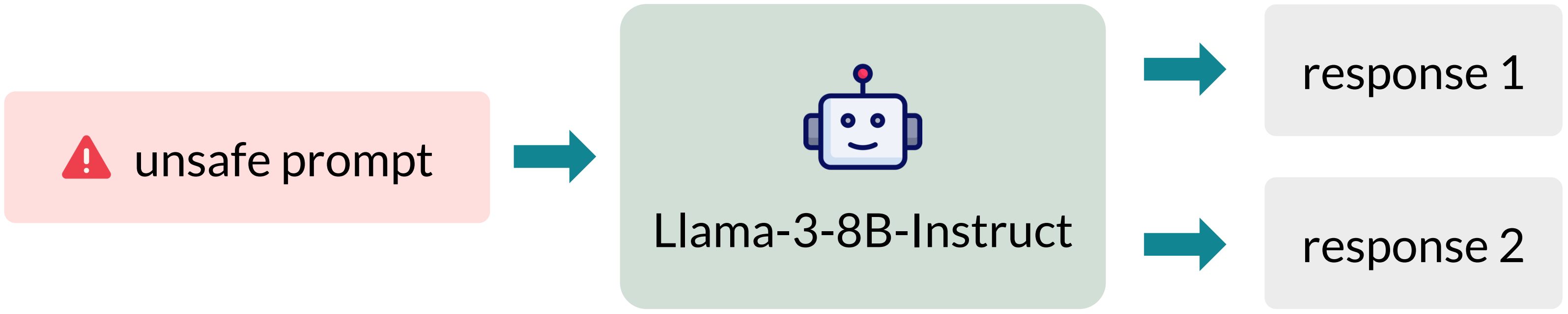
Preference Dataset: Unsafe prompts from SORRY-Bench (Xie et al. 2024)



Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a (moderately aligned) language model to refuse unsafe prompts via DPO

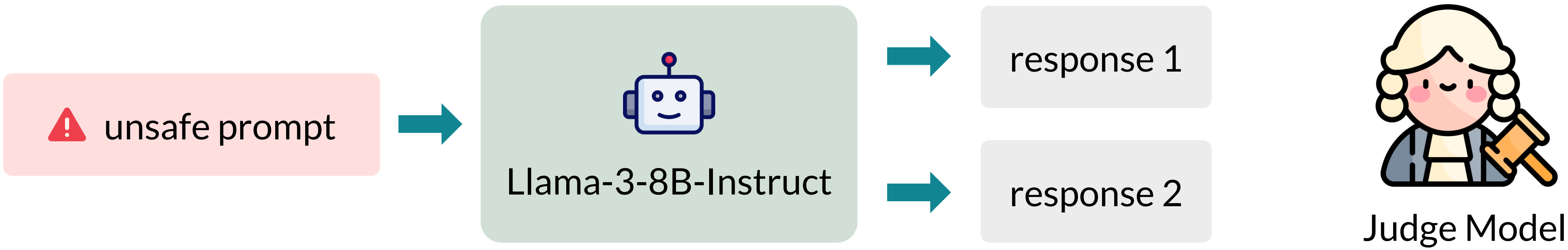
Preference Dataset: Unsafe prompts from SORRY-Bench (Xie et al. 2024)



Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a (moderately aligned) language model to refuse unsafe prompts via DPO

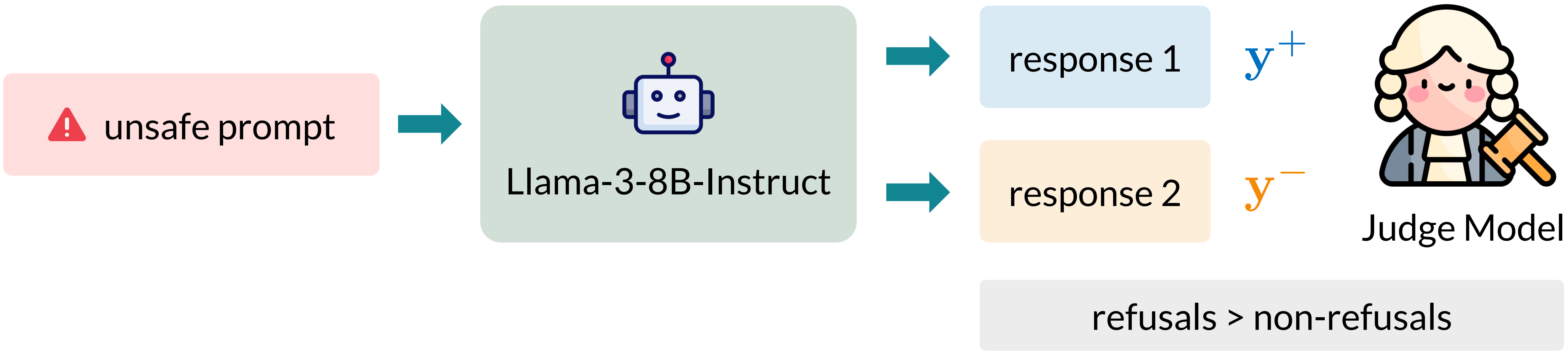
Preference Dataset: Unsafe prompts from SORRY-Bench (Xie et al. 2024)



Likelihood Displacement Can Cause Unintentional Unalignment

Setting: Train a (moderately aligned) language model to refuse unsafe prompts via DPO

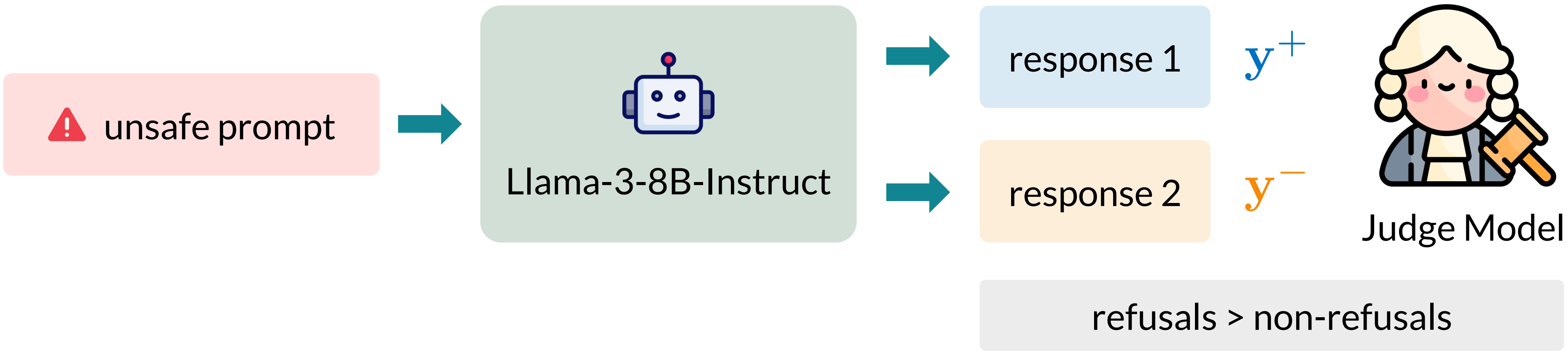
Preference Dataset: Unsafe prompts from SORRY-Bench (Xie et al. 2024)



Likelihood Displacement Can Cause Unintentional Unalignment

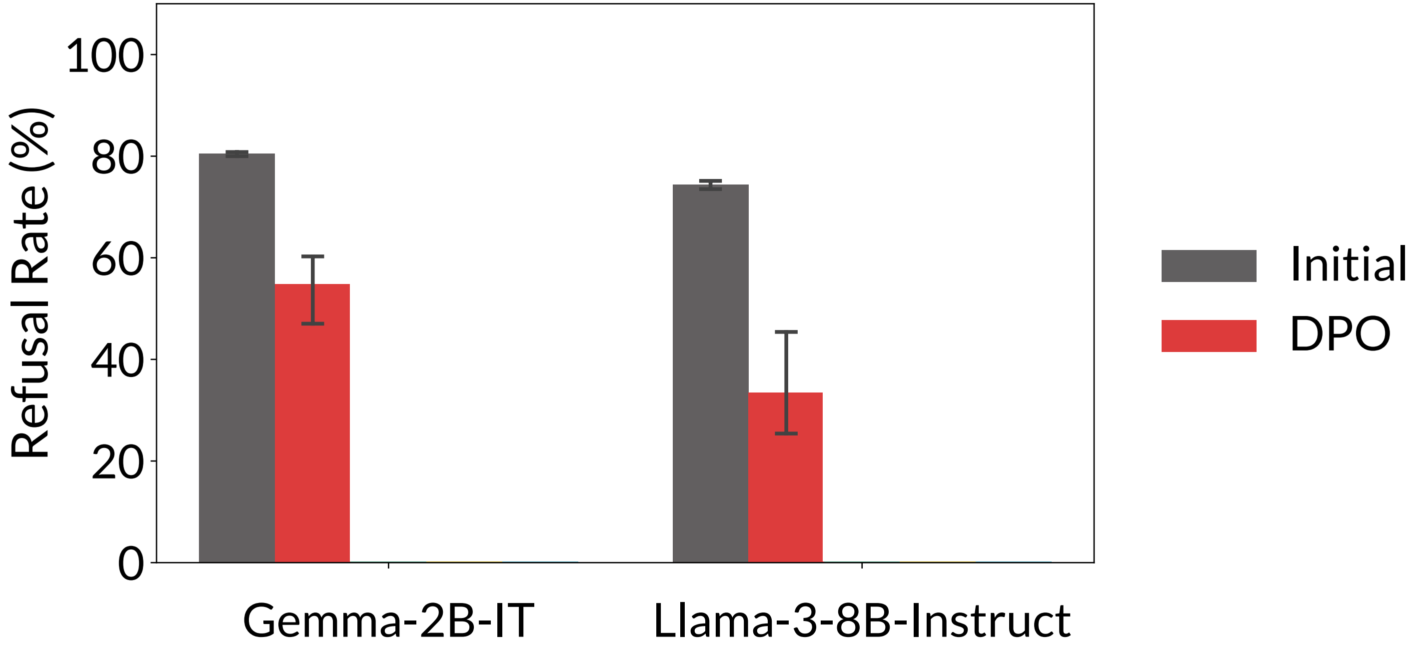
Setting: Train a (moderately aligned) language model to refuse unsafe prompts via DPO

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



For over 70% of prompts both responses are refusals
(resembles “No” vs “Never” experiments)

Likelihood Displacement Can Cause Unintentional Unalignment



Likelihood displacement leads to unintentional unalignment!

Main Contributions

- Likelihood displacement can be catastrophic and lead to surprising failures in alignment
- Theory: Likelihood displacement is driven by the model’s embedding geometry
- Mitigating likelihood displacement via data filtering

Theoretical Analysis of Likelihood Displacement: Approach

Theoretical Analysis of Likelihood Displacement: Approach

Goal: Characterize how $\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ changes during training

response

prompt

Theoretical Analysis of Likelihood Displacement: Approach

Goal: Characterize how $\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ changes during training

response prompt

$\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ is determined by:

Theoretical Analysis of Likelihood Displacement: Approach

Goal: Characterize how $\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ changes during training

response prompt

$\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ is determined by:

1 hidden embeddings $\mathbf{h}_{\mathbf{x},\mathbf{z}_{<1}}, \ldots, \mathbf{h}_{\mathbf{x},\mathbf{z}_{<|\mathbf{z}|}}$

Theoretical Analysis of Likelihood Displacement: Approach

Goal: Characterize how $\ln \pi_\theta(\underbrace{\mathbf{z}}_{\text{response}}|\underbrace{\mathbf{x}}_{\text{prompt}})$ changes during training

$\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ is determined by:

1 hidden embeddings $\mathbf{h}_{\mathbf{x},\mathbf{z}_{<1}}, \ldots, \mathbf{h}_{\mathbf{x},\mathbf{z}_{<|\mathbf{z}|}}$

2 token unembeddings matrix $\mathbf{W}$

Theoretical Analysis of Likelihood Displacement: Approach

Goal: Characterize how $\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ changes during training

response prompt

$\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ is determined by:

$$\underbrace{\boxed{\text{1 hidden embeddings } \mathbf{h}_{\mathbf{x},\mathbf{z}_{<1}}, \ldots, \mathbf{h}_{\mathbf{x},\mathbf{z}_{<|\mathbf{z}|}}} \quad \boxed{\text{2 token unembeddings matrix } \mathbf{W}}}_{\text{We track their evolution during training}}$$

Theoretical Analysis of Likelihood Displacement: Approach

Goal: Characterize how $\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ changes during training

response prompt

$\ln \pi_\theta(\mathbf{z}|\mathbf{x})$ is determined by:

$$\underbrace{\boxed{\text{1 hidden embeddings } \mathbf{h}_{\mathbf{x},\mathbf{z}_{<1}}, \ldots, \mathbf{h}_{\mathbf{x},\mathbf{z}_{<|\mathbf{z}|}}} \quad \boxed{\text{2 token unembeddings matrix } \mathbf{W}}}_{\text{We track their evolution during training}}$$

Assumption: For simplicity, consider hidden embeddings as trainable parameters
(Suanshi et al. 2021, Zhu et al. 2021, Mixon et al. 2022, Ji et al. 2022, Tirer et al. 2023)

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Theorem: When does likelihood displacement occur?

At any training step, $\ln \pi_\theta(\mathbf{y}^+|\mathbf{x})$ decreases when the following are large:

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Theorem: When does likelihood displacement occur?

At any training step, $\ln \pi_\theta(\mathbf{y}^+|\mathbf{x})$ decreases when the following are large:

1. $\langle \mathbf{W}_{\mathbf{y}^+}, \mathbf{W}_{\mathbf{y}^-} \rangle$ — Intuition: similar preferences cause likelihood displacement

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Theorem: When does likelihood displacement occur?

At any training step, $\ln \pi_\theta (\mathbf{y}^+|\mathbf{x})$ decreases when the following are large:

1. $\langle \mathbf{W}_{\mathbf{y}^+}, \mathbf{W}_{\mathbf{y}^-} \rangle$ — Intuition: similar preferences cause likelihood displacement

2. $\langle \mathbf{W}_{\mathbf{z}}, \mathbf{W}_{\mathbf{y}^+} - \mathbf{W}_{\mathbf{y}^-} \rangle$ for tokens $\mathbf{z} \neq \mathbf{y}^+, \mathbf{y}^-$

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Theorem: Where does the probability mass go?

The log probability change of $\mathbf{z}$ is proportional to: $\langle \mathbf{W}_{\mathbf{z}}, \mathbf{W}_{\mathbf{y}^+} - \mathbf{W}_{\mathbf{y}^-} \rangle$

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Theorem: Where does the probability mass go?

The log probability change of $\mathbf{z}$ is proportional to: $\langle \mathbf{W}_{\mathbf{z}}, \mathbf{W}_{\mathbf{y}^+} - \mathbf{W}_{\mathbf{y}^-} \rangle$

Empirical Observation: $\mathbf{W}_{\mathbf{y}^+} - \mathbf{W}_{\mathbf{y}^-}$ often has a large component orthogonal to $\mathbf{W}_{\mathbf{y}^+}$

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Theorem: Where does the probability mass go?

The log probability change of $\mathbf{z}$ is proportional to: $\langle \mathbf{W}_{\mathbf{z}}, \mathbf{W}_{\mathbf{y}^+} - \mathbf{W}_{\mathbf{y}^-} \rangle$

Empirical Observation: $\mathbf{W}_{\mathbf{y}^+} - \mathbf{W}_{\mathbf{y}^-}$ often has a large component orthogonal to $\mathbf{W}_{\mathbf{y}^+}$

Token unembeddings encode semantics

(e.g. Mikolov et al. 2013, Park et al. 2024)

Single Token Responses: Role of Token Unembedding Geometry

Suppose that $\mathbf{y}^+$ and $\mathbf{y}^-$ consist of a single token

Theorem: Where does the probability mass go?

The log probability change of $\mathbf{z}$ is proportional to: $\langle \mathbf{W}_{\mathbf{z}}, \mathbf{W}_{\mathbf{y}^+} - \mathbf{W}_{\mathbf{y}^-} \rangle$

Empirical Observation: $\mathbf{W}_{\mathbf{y}^+} - \mathbf{W}_{\mathbf{y}^-}$ often has a large component orthogonal to $\mathbf{W}_{\mathbf{y}^+}$

Token unembeddings encode semantics
(e.g. Mikolov et al. 2013, Park et al. 2024)

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}$$

Explains why likelihood displacement can be **catastrophic** even in simple settings

Multiple Token Responses: Role of Hidden Embedding Geometry

Consider the typical case where $\mathbf{y}^+$ and $\mathbf{y}^-$ are sequences

Multiple Token Responses: Role of Hidden Embedding Geometry

Consider the typical case where $\mathbf{y}^+$ and $\mathbf{y}^-$ are sequences

Theorem: When does likelihood displacement occur?

At any training step, in addition to the dependence on token unembeddings, $\ln \pi_\theta \left(\mathbf{y}^+ | \mathbf{x} \right)$ decreases more the larger the following term is:

Multiple Token Responses: Role of Hidden Embedding Geometry

Consider the typical case where $\mathbf{y}^+$ and $\mathbf{y}^-$ are sequences

Theorem: When does likelihood displacement occur?

At any training step, in addition to the dependence on token unembeddings, $\ln \pi_\theta\left(\mathbf{y}^+|\mathbf{x}\right)$ decreases more the larger the following term is:

$$\sum_{k=1}^{|\mathbf{y}^+|}\sum_{k'=1}^{|\mathbf{y}^-|} \alpha^-_{k,k'} \cdot \underbrace{\left\langle \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^-_{<k'}} \right\rangle}_{\text{preferred-dispreferred alignment}} - \sum_{k=1}^{|\mathbf{y}^+|}\sum_{k'=1}^{|\mathbf{y}^+|} \alpha^+_{k,k'} \cdot \underbrace{\left\langle \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k'}} \right\rangle}_{\text{preferred-preferred alignment}}$$

Multiple Token Responses: Role of Hidden Embedding Geometry

Consider the typical case where $\mathbf{y}^+$ and $\mathbf{y}^-$ are sequences

Theorem: When does likelihood displacement occur?

At any training step, in addition to the dependence on token unembeddings, $\ln \pi_\theta\left(\mathbf{y}^+|\mathbf{x}\right)$ decreases more the larger the following term is:

$$\sum_{k=1}^{|\mathbf{y}^+|} \sum_{k'=1}^{|\mathbf{y}^-|} \alpha^-_{k,k'} \cdot \underbrace{\left\langle \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^-_{<k'}} \right\rangle}_{\text{preferred-dispreferred alignment}} - \sum_{k=1}^{|\mathbf{y}^+|} \sum_{k'=1}^{|\mathbf{y}^+|} \alpha^+_{k,k'} \cdot \underbrace{\left\langle \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k'}} \right\rangle}_{\text{preferred-preferred alignment}}$$

$\alpha^-_{k,k'}, \alpha^+_{k,k'} \in [-2, 2]$ are determined by the model's next-token distributions

Centered Hidden Embedding Similarity (CHES) Score

Empirical Observation: $\alpha^{-}_{k,k'}, \alpha^{+}_{k,k'}$ coefficients are mostly positive

Centered Hidden Embedding Similarity (CHES) Score

Empirical Observation: $\alpha^{-}_{k,k'}, \alpha^{+}_{k,k'}$ coefficients are mostly positive

Accordingly, setting these coefficients to 1 leads to:

Centered Hidden Embedding Similarity (CHES) Score

Empirical Observation: $\alpha^-_{k,k'}, \alpha^+_{k,k'}$ coefficients are mostly positive

Accordingly, setting these coefficients to 1 leads to:

Definition: Centered Hidden Embedding Similarity (CHES) Score

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

*The CHES score is model-dependent

Centered Hidden Embedding Similarity (CHES) Score

Empirical Observation: $\alpha^{-}_{k,k'}, \alpha^{+}_{k,k'}$ coefficients are mostly positive

Accordingly, setting these coefficients to 1 leads to:

Definition: Centered Hidden Embedding Similarity (CHES) Score

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

*The CHES score is model-dependent

Our theory indicates that a higher CHES score leads to more likelihood displacement

Main Contributions

Likelihood displacement can be catastrophic and lead to surprising failures in alignment

Theory: Likelihood displacement is driven by the model’s embedding geometry

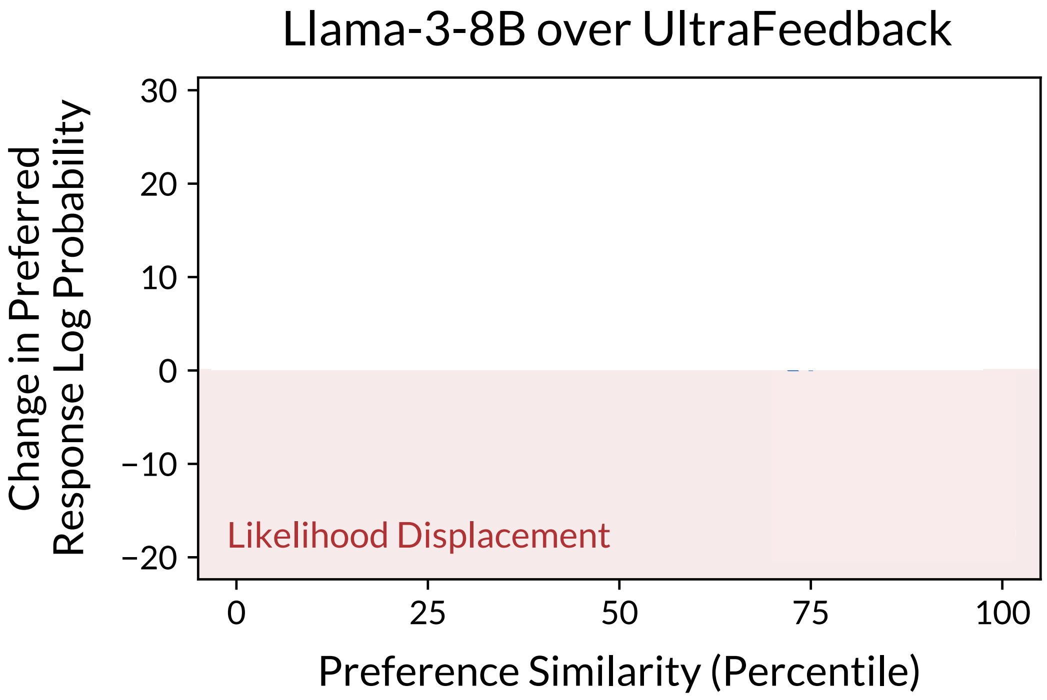
Mitigating likelihood displacement via data filtering

Identifying Sources of Likelihood Displacement

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

Identifying Sources of Likelihood Displacement

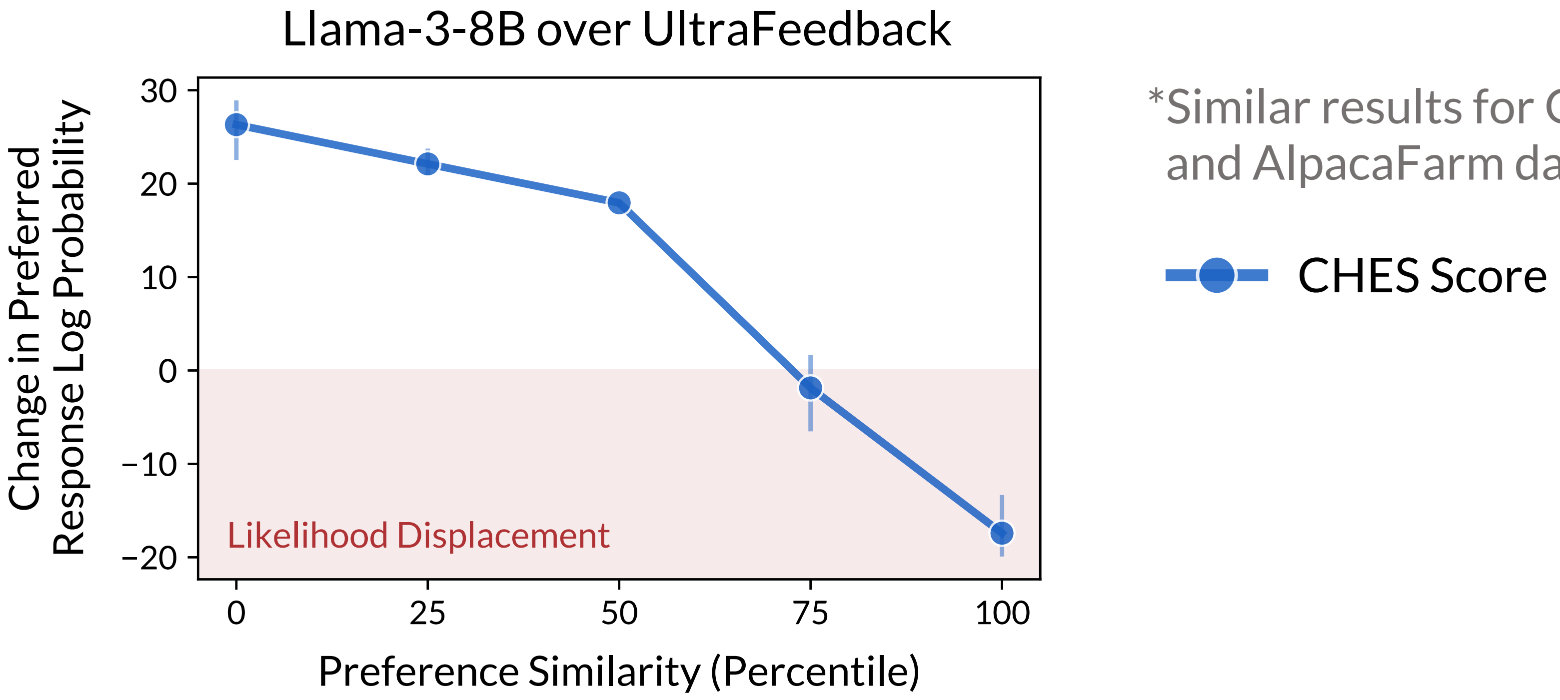
Q: How indicative is the CHES score of likelihood displacement?



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

Identifying Sources of Likelihood Displacement

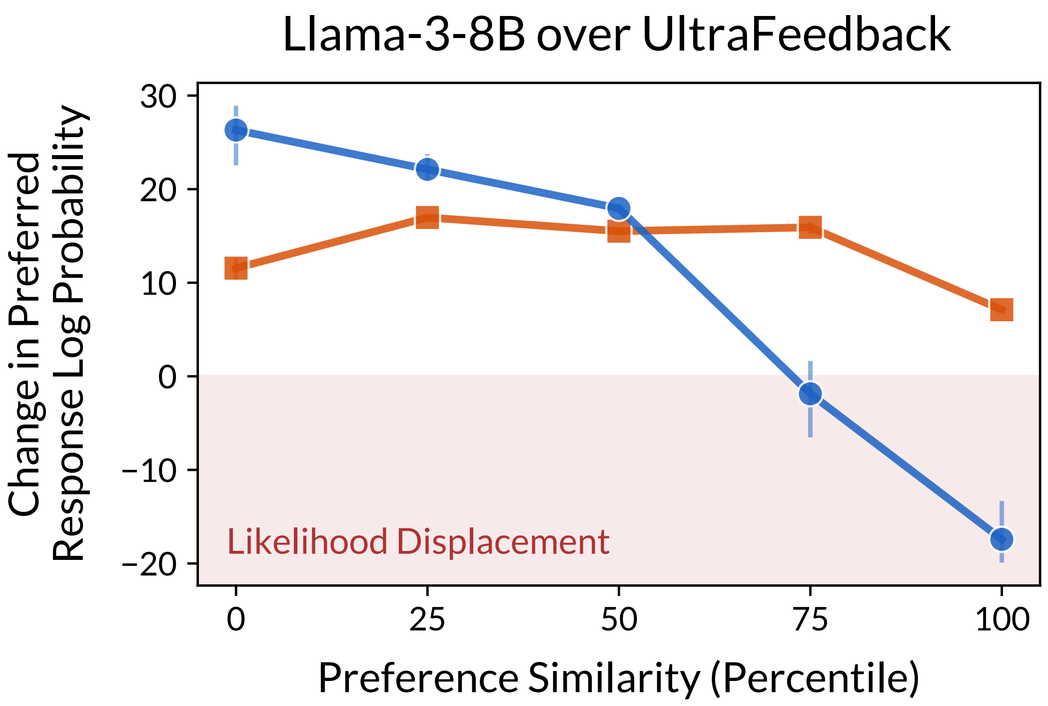
Q: How indicative is the CHES score of likelihood displacement?



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

Identifying Sources of Likelihood Displacement

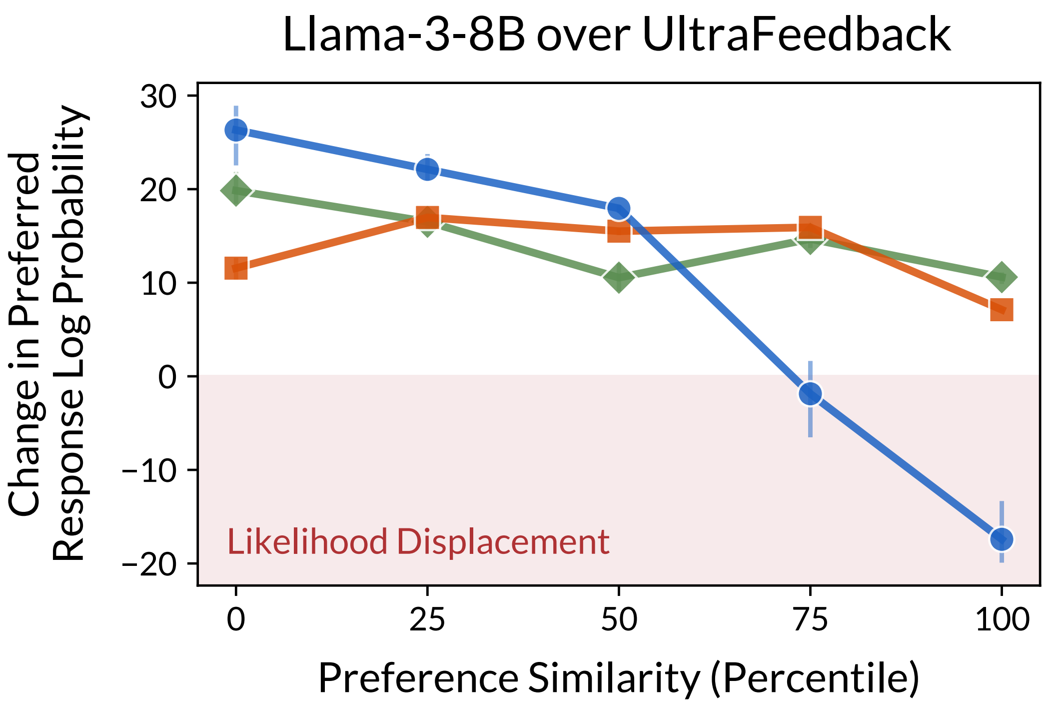
Q: How indicative is the CHES score of likelihood displacement?



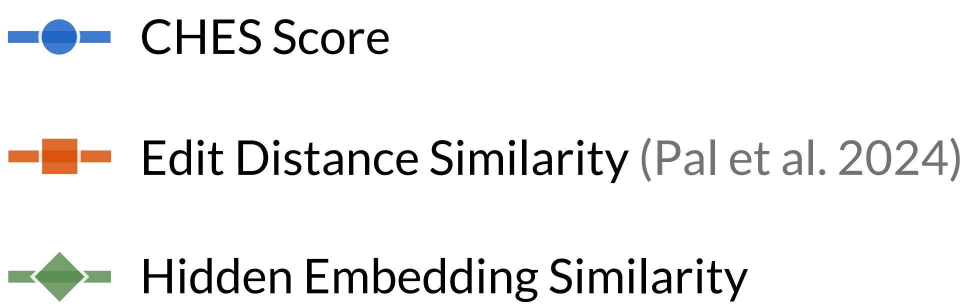
*Similar results for OLMo-1B, Gemma-2B models and AlpacaFarm dataset

Identifying Sources of Likelihood Displacement

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

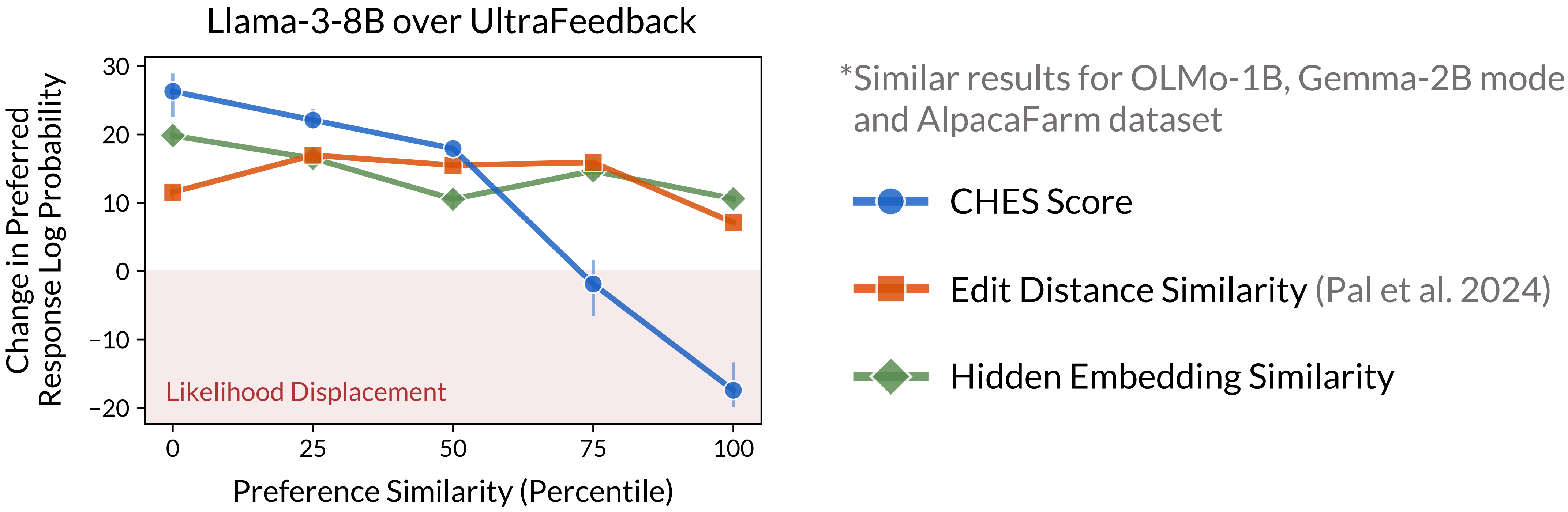


*Similar results for OLMo-1B, Gemma-2B models and AlpacaFarm dataset



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

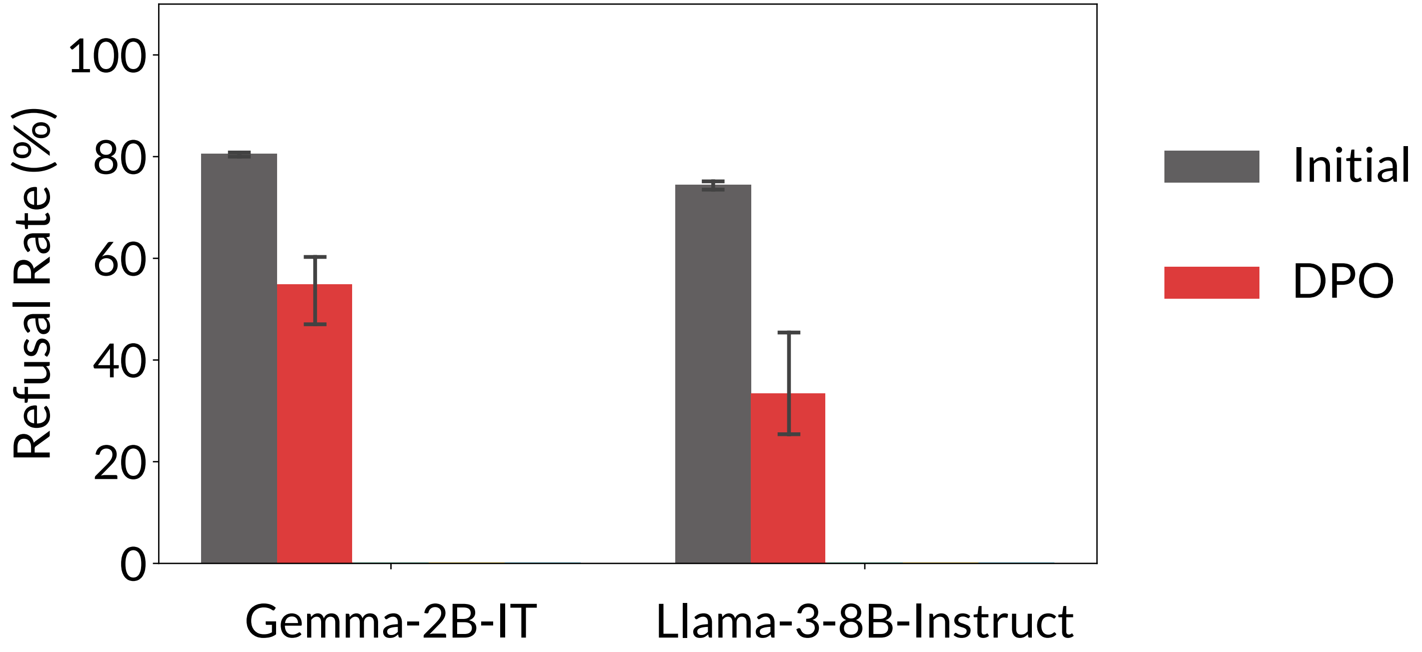
CHES score identifies training samples causing likelihood displacement, whereas alternative measures do not

Data Filtering via CHES Score Mitigates Unintentional Unalignment

Recall: Unintentional unalignment due to likelihood displacement experiments

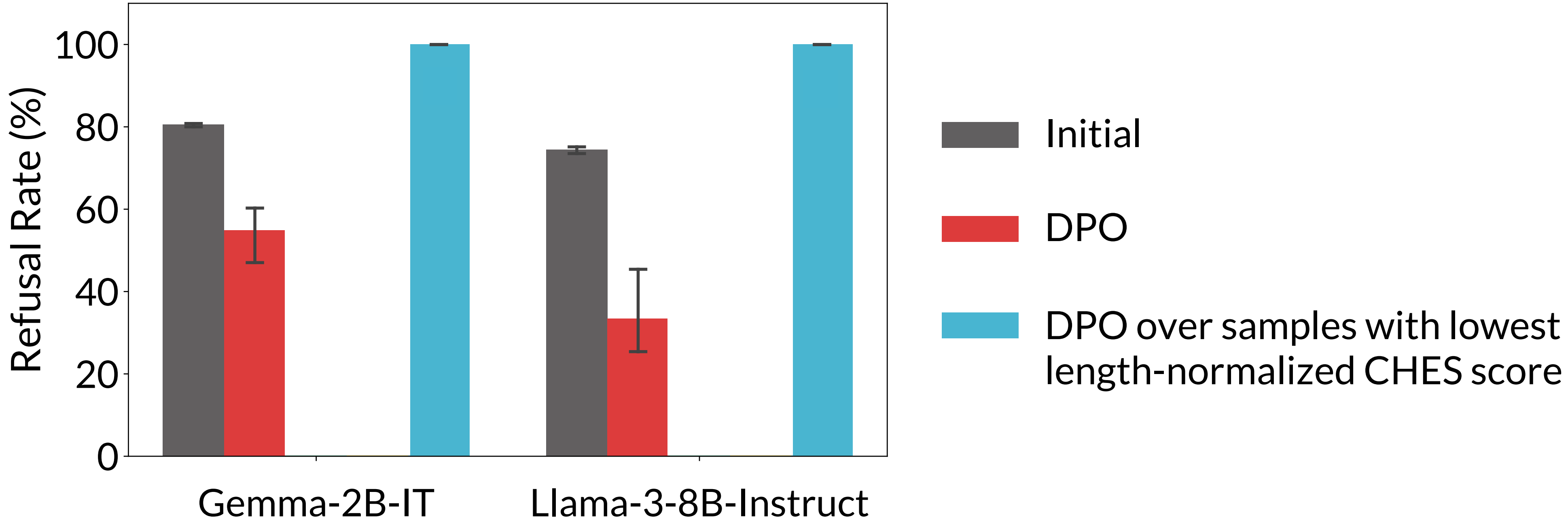
Data Filtering via CHES Score Mitigates Unintentional Unalignment

Recall: Unintentional unalignment due to likelihood displacement experiments



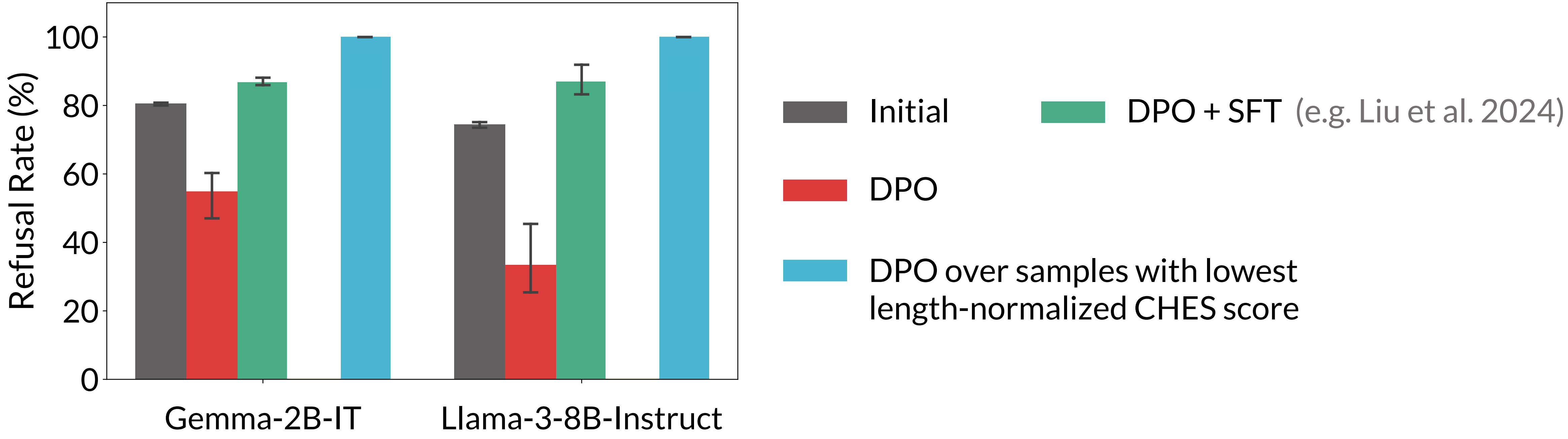
Data Filtering via CHES Score Mitigates Unintentional Unalignment

Recall: Unintentional unalignment due to likelihood displacement experiments



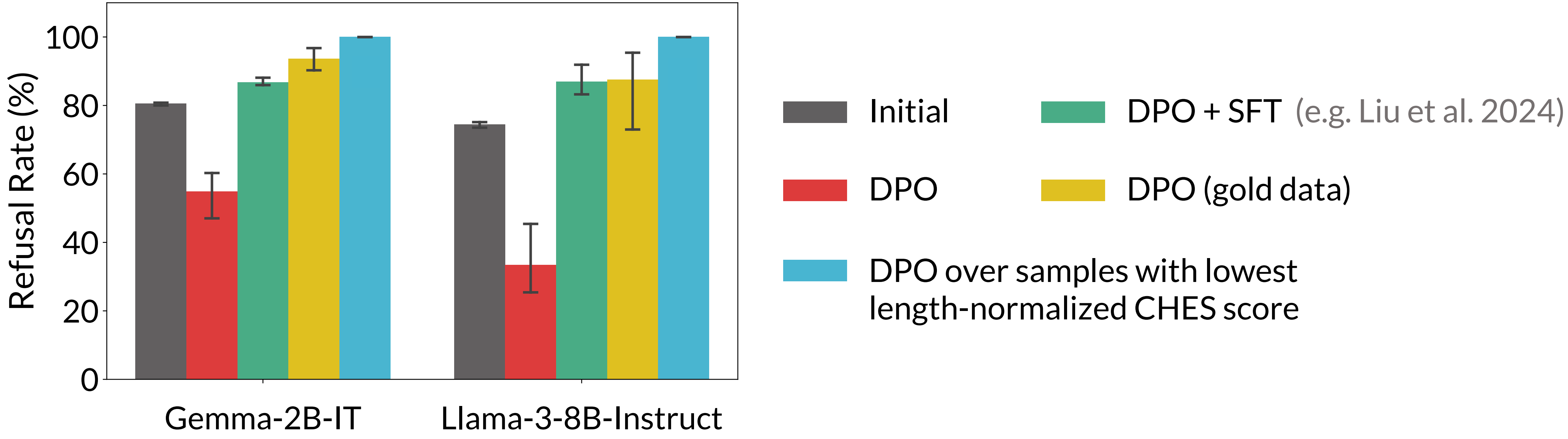
Data Filtering via CHES Score Mitigates Unintentional Unalignment

Recall: Unintentional unalignment due to likelihood displacement experiments



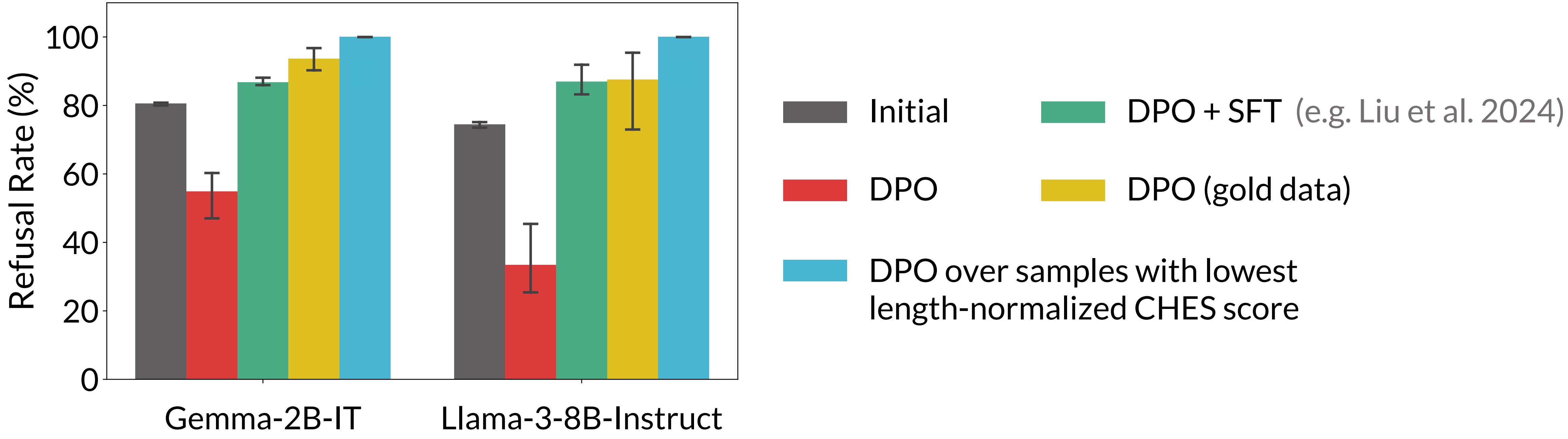
Data Filtering via CHES Score Mitigates Unintentional Unalignment

Recall: Unintentional unalignment due to likelihood displacement experiments



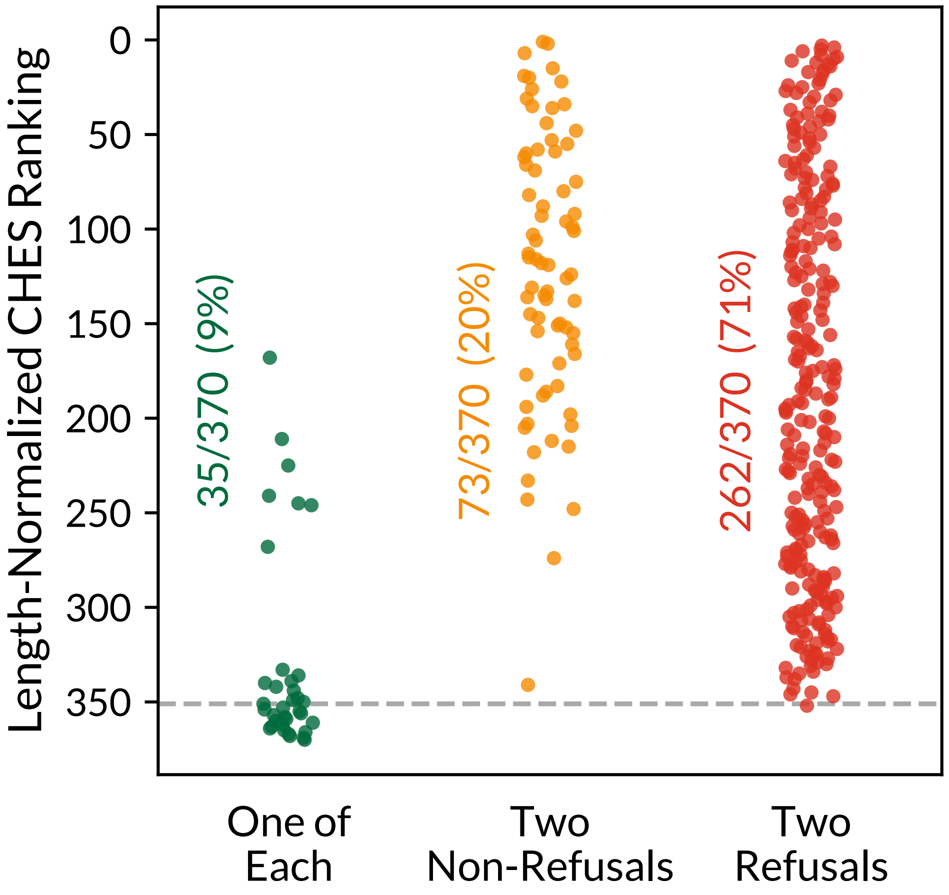
Data Filtering via CHES Score Mitigates Unintentional Unalignment

Recall: Unintentional unalignment due to likelihood displacement experiments



Removing samples with high CHES scores mitigates unintentional unalignment, and goes beyond adding an SFT term to the loss

Which Samples Have a High CHES Score?



CHES score ranking falls in line with intuition: Samples with **two refusal** or **two non-refusal** responses tend to have a higher score than samples with **one of each**

Conclusion

Conclusion

Likelihood displacement can be catastrophic and cause **unintentional unlignment**

Conclusion

Likelihood displacement can be catastrophic and cause **unintentional unlignment**

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

Conclusion

- Likelihood displacement can be catastrophic and cause **unintentional unlignment**
- Theory & Experiments: Samples with **high CHES scores lead to likelihood displacement**
- **Filtering out samples with high CHES score** can mitigate unintentional unalignment

Conclusion

Likelihood displacement can be catastrophic and cause **unintentional unlignment**

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

Filtering out samples with high CHES score can mitigate unintentional unalignment

Our work highlights the importance of curating data with sufficiently distinct preferences, for which the CHES score may prove valuable

Outlook

Fundamentals of Language Model Alignment

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF Ouyang et al. 2022	RAFT Dong et al. 2023	IPO Azar et al. 2023	REBEL Gao et al. 2024	KTO Ethayarajh et al. 2024
RLAIF Bai et al. 2022	DPO Rafailov et al. 2023	SLiC-HF Zhao et al. 2023	SimPO Meng et al. 2024	...

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF Ouyang et al. 2022	RAFT Dong et al. 2023	IPO Azar et al. 2023	REBEL Gao et al. 2024	KTO Ethayarajh et al. 2024
RLAIF Bai et al. 2022	DPO Rafailov et al. 2023	SLiC-HF Zhao et al. 2023	SimPO Meng et al. 2024	...

Limited understanding of basic questions (e.g. loss landscape, optimization, generalization)

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF Ouyang et al. 2022	RAFT Dong et al. 2023	IPO Azar et al. 2023	REBEL Gao et al. 2024	KTO Ethayarajh et al. 2024
RLAIF Bai et al. 2022	DPO Rafailov et al. 2023	SLiC-HF Zhao et al. 2023	SimPO Meng et al. 2024	...

Limited understanding of basic questions (e.g. loss landscape, optimization, generalization)

Theory (mathematical or empirical) may be necessary for efficient and reliable alignment

Fundamentals of Language Model Alignment

There are countless methods for aligning language models

RLHF Ouyang et al. 2022	RAFT Dong et al. 2023	IPO Azar et al. 2023	REBEL Gao et al. 2024	KTO Ethayarajh et al. 2024
RLAIF Bai et al. 2022	DPO Rafailov et al. 2023	SLiC-HF Zhao et al. 2023	SimPO Meng et al. 2024	...

Limited understanding of basic questions (e.g. loss landscape, optimization, generalization)

Theory (mathematical or empirical) may be necessary for efficient and reliable alignment

Thank You!

Work supported in part by the Zuckerman STEM Leadership Program