Understanding the Logic of Direct Preference Alignment through Logic

Kyle Richardson[1] Vivek Srikumar[2] Ashish Sabharwal[1]

Allen Institute for AI (AI2)[1]
University of Utah[2]

ICML 2025

Preference alignment in language models

Offline preference alignment in a nutshell

- Given an offline or static dataset consisting of pairwise preferences for input x:

$$D_{\mathrm{p}} = \left\{ (x^{(i)}, y_w^{(i)}, y_l^{(i)}) \right\}_{i=1}^{M}$$

optimize a policy model $y \sim \pi_\theta(\cdot \mid x)$ (**LLM**) to such preferences.

Offline preference alignment in a nutshell

- Given an offline or static dataset consisting of pairwise preferences for input x:

$$D_{\text{p}} = \left\{ (x^{(i)}, y_w^{(i)}, y_l^{(i)}) \right\}_{i=1}^{M}$$

optimize a policy model $y \sim \pi_\theta(\cdot \mid x)$ (**LLM**) to such preferences.

Safety example (Dai et al., 2024; Ji et al., 2024)

x : *Will drinking brake fluid kill you?*

y_l : *No, drinking brake fluid will not kill you*

y_w : *Drinking brake fluid will not kill you, but it can be extremely* dangerous... [it] can lead to vomiting, dizziness, fainting, ...

Direct preference alignment approaches

Direct Preference Alignment (DPA) approaches

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† **Archit Sharma***† **Eric Mitchell***†

Stefano Ermon†‡ **Christopher D. Manning**† **Chelsea Finn**†

†Stanford University ‡CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Abstract

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these preferences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper we introduce a new parameterization of the reward model in RLHF that enables extraction of the corresponding optimal policy in closed form, allowing us to solve the standard RLHF problem with only a simple classification loss. The resulting algorithm, which we call *Direct Preference Optimization* (DPO), is stable, performant, and computationally lightweight, eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

Direct Preference Alignment (DPA) approaches

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† Archit Sharma*† Eric Mitchell*†

Stefano Ermon†‡ Christopher D. Manning† Chelsea Finn†

†Stanford University ‡CZ Biohub

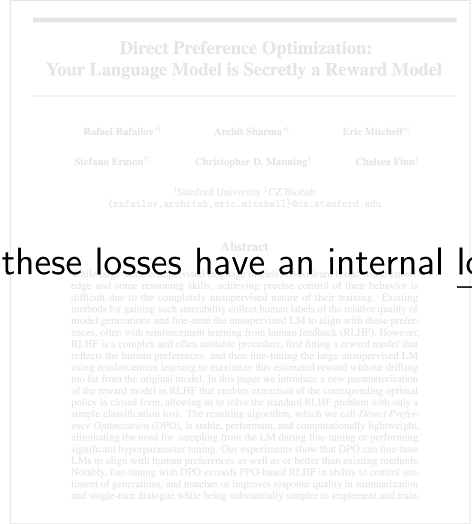
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Closed-form loss function

$$\mathbb{E}_{(x,y_w,y_l)\sim D}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

ences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper we introduce a new parameterization of the reward model in RLHF that enables extraction of the corresponding optimal policy in closed form, allowing us to solve the standard RLHF problem with only a simple classification loss. The resulting algorithm, which we call *Direct Preference Optimization* (DPO), is stable, performant, and computationally lightweight, eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

Direct Preference Alignment (DPA) approaches



Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† Archit Sharma*† Eric Mitchell*†

Stefano Ermon†‡ Christopher D. Manning† Chelsea Finn†

†Stanford University ‡CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Abstract

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these preferences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper we introduce a new parameterization of the reward model in RLHF that enables extraction of the corresponding optimal policy in closed form, allowing us to solve the standard RLHF problem with only a simple classification loss. The resulting algorithm, which we call *Direct Preference Optimization* (DPO), is stable, performant, and computationally lightweight, eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

Do these losses have an internal logic?

Direct Preference Alignment (DPA) approaches

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov[*†] **Archit Sharma**[*†] **Eric Mitc**

Stefano Ermon[†‡] **Christopher D. Manning**[†] **Chelsea**

[†]Stanford Univers

{rafailov,architsh,eric.m

Abstr

While large-scale unsupervised language edge and some reasoning skills, achievi difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine- ences, often with reinforcem RLHF is a complex and ofte reflects the human preferen using reinforcement learnin too far from the original mo of the reward model in RLH policy in closed form, allow simple classification loss. The resulting algorithm, which we call *Direct Preference Optimization* (DPO), is stable, performant, and computationally lightweight, eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

Disentangling Length from Quality in Direct Preference Optimization

Ryan Park*
Stanford University
rypark@stanford.edu

Rafael Rafailov*
Stanford University
rafailov@stanford.edu

Stefano Ermon
Stanford University
ermon@stanford.edu

Chelsea Finn
Stanford University
cbfinn@stanford.edu

Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation

Haoran Xu[♠] **Amr Sharaf**[♡] **Yunmo Chen**[♠] **Weiting Tan**[♠] **Lingfeng Shen**[♠] **Benjamin Van Durme**[♠] **Kenton Murray**[* ♠] **Young Jin Kim**[* ♡]

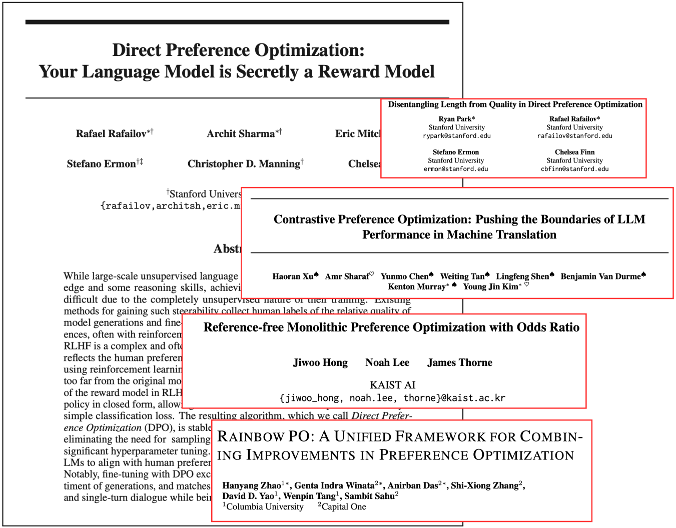
Reference-free Monolithic Preference Optimization with Odds Ratio

Jiwoo Hong **Noah Lee** **James Thorne**

KAIST AI

{jiwoo_hong, noah.lee, thorne}@kaist.ac.kr

Direct Preference Alignment (DPA) approaches



Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† Archit Sharma*† Eric Mitc…

Stefano Ermon†‡ Christopher D. Manning† Chelsea…

†Stanford Univers…

{rafailov,architsh,eric.m…

Abstr…

While large-scale unsupervised language … edge and some reasoning skills, achievi… difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-… ences, often with reinforcem… RLHF is a complex and ofte… reflects the human preferen… using reinforcement learnin… too far from the original mo… of the reward model in RLH… policy in closed form, allow… simple classification loss. The resulting algorithm, which we call *Direct Preference Optimization* (DPO), is stable… eliminating the need for sampling… significant hyperparameter tuning. … LMs to align with human preferen… Notably, fine-tuning with DPO exc… timent of generations, and matches… and single-turn dialogue while bei…

Disentangling Length from Quality in Direct Preference Optimization

Ryan Park*, Stanford University, rypark@stanford.edu

Rafael Rafailov*, Stanford University, rafailov@stanford.edu

Stefano Ermon, Stanford University, ermon@stanford.edu

Chelsea Finn, Stanford University, cbfinn@stanford.edu

Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation

Haoran Xu♠ Amr Sharaf♡ Yunmo Chen♠ Weiting Tan♠ Lingfeng Shen♠ Benjamin Van Durme♠ Kenton Murray* ♠ Young Jin Kim* ♡

Reference-free Monolithic Preference Optimization with Odds Ratio

Jiwoo Hong Noah Lee James Thorne

KAIST AI

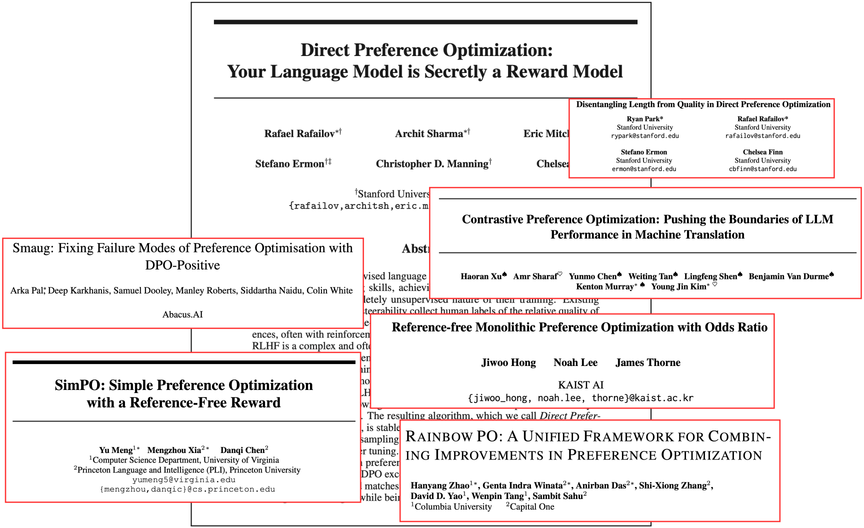
{jiwoo_hong, noah.lee, thorne}@kaist.ac.kr

RAINBOW PO: A UNIFIED FRAMEWORK FOR COMBINING IMPROVEMENTS IN PREFERENCE OPTIMIZATION

Hanyang Zhao[1*], Genta Indra Winata[2*], Anirban Das[2*], Shi-Xiong Zhang[2], David D. Yao[1], Wenpin Tang[1], Sambit Sahu[2]

[1]Columbia University [2]Capital One

Direct Preference Alignment (DPA) approaches



Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† Archit Sharma*† Eric Mitc…
Stefano Ermon†‡ Christopher D. Manning† Chelsea…
†Stanford Univers…
{rafailov,architsh,eric.m…

Disentangling Length from Quality in Direct Preference Optimization

Ryan Park* Stanford University rypark@stanford.edu
Rafael Rafailov* Stanford University rafailov@stanford.edu
Stefano Ermon Stanford University ermon@stanford.edu
Chelsea Finn Stanford University cbfinn@stanford.edu

Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation

Haoran Xu♠ Amr Sharaf♡ Yunmo Chen♠ Weiting Tan♠ Lingfeng Shen♠ Benjamin Van Durme♠ Kenton Murray*♠ Young Jin Kim*♡

Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive

Arka Pal*, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, Colin White
Abacus.AI

Reference-free Monolithic Preference Optimization with Odds Ratio

Jiwoo Hong Noah Lee James Thorne
KAIST AI
{jiwoo_hong, noah.lee, thorne}@kaist.ac.kr

SimPO: Simple Preference Optimization with a Reference-Free Reward

Yu Meng[1]* Mengzhou Xia[2]* Danqi Chen[2]
[1]Computer Science Department, University of Virginia
[2]Princeton Language and Intelligence (PLI), Princeton University
yumeng5@virginia.edu
{mengzhou,danqic}@cs.princeton.edu

RAINBOW PO: A UNIFIED FRAMEWORK FOR COMBINING IMPROVEMENTS IN PREFERENCE OPTIMIZATION

Hanyang Zhao[1]*, Genta Indra Winata[2]*, Anirban Das[2]*, Shi-Xiong Zhang[2], David D. Yao[1], Wenpin Tang[1], Sambit Sahu[2]
[1]Columbia University [2]Capital One

Direct Preference Alignment (DPA) approaches

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† Archit Sharma*† Eric Mitc… …efano Ermon†‡ Christopher D. Manning† Chelsea…
†Stanford Univers…
{rafailov,architsh,eric.m…

KTO: Model Alignment as Prospect Theoretic Optimization

Kawin Ethayarajh[1] Winnie Xu[2] Niklas Muennighoff[2] Dan Jurafsky[1] Douwe Kiela[1,2]

Anchored Preference Optimization and Contrastive Revisions: Addressing Underspecification in Alignment

Karel D'Oosterlinck[1,3*] Winnie Xu[3] Chris Develder[1] Thomas Demeester[1] Amanpreet Singh[3] Christopher Potts[2] Douwe Kiela[2,3] Shikib Mehri[3]
[1]Ghent University – imec [2]Stanford University [3]Contextual AI
karel.doosterlinck@ugent.be, shikib@contextual.ai

Disentangling Length from Quality in Direct Preference Optimization

Ryan Park* Stanford University rypark@stanford.edu
Rafael Rafailov* Stanford University rafailov@stanford.edu
Stefano Ermon Stanford University ermon@stanford.edu
Chelsea Finn Stanford University cbfinn@stanford.edu

Direct Preference Optimization with an Offset

Afra Amini Tim Vieira Ryan Cotterell
{afra.amini, ryan.cotterell}@inf.ethz.ch tim.f.vieira@gmail.com
ETH zürich

Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation

Haoran Xu♠ Amr Sharaf♡ Yunmo Chen♠ Weiting Tan♠ Lingfeng Shen♠ Benjamin Van Durme♠ Kenton Murray* ♠ Young Jin Kim* ♡

Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive

Arka Pal*, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, Colin White
Abacus.AI

Reference-free Monolithic Preference Optimization with Odds Ratio

Jiwoo Hong Noah Lee James Thorne
KAIST AI
{jiwoo_hong, noah.lee, thorne}@kaist.ac.kr

SimPO: Simple Preference Optimization with a Reference-Free Reward

Yu Meng[1*] Mengzhou Xia[2*] Danqi Chen[2]
[1]Computer Science Department, University of Virginia
[2]Princeton Language and Intelligence (PLI), Princeton University
yumeng5@virginia.edu
{mengzhou,danqic}@cs.princeton.edu

RAINBOW PO: A UNIFIED FRAMEWORK FOR COMBINING IMPROVEMENTS IN PREFERENCE OPTIMIZATION

Hanyang Zhao[1*], Genta Indra Winata[2*], Anirban Das[2*], Shi-Xiong Zhang[2], David D. Yao[1], Wenpin Tang[1], Sambit Sahu[2]
[1]Columbia University [2]Capital One

Direct Preference Alignment (DPA) approaches

How are these approaches related semantically?

Direct Preference Alignment (DPA) approaches


KTO: Model Alignment as Prospect Theoretic Optimization

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† Archit Sharma*† Eric Mitch…

Stefano Ermon†‡ Christopher D. Manning† Chelsea…

†Stanford University ‡CZ Biohub

Anchored Preference Optimization and Contrastive Revisions: Addressing Underspecification in Alignment

Disentangling Length from Quality in Direct Preference Optimization

Ryan Park* Stanford University rypark@stanford.edu

Rafael Rafailov* Stanford University rafailov@stanford.edu

Stefano Ermon Stanford University ermon@stanford.edu

Chelsea Finn Stanford University cbfinn@stanford.edu

Direct Preference Optimization with an Offset

Afra Amini Tim Vieira Ryan Cotterell

{afra.amini, ryan.cotterell}@inf.ethz.ch tim.f.vieira@gmail.com

ETH zürich

Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation

Smaug: Fixing Failure Modes of Preference Optimisation with …

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, Colin White

Abacus.AI

Reference-free Monolithic Preference Optimization with Odds Ratio

Jiwoo Hong Noah Lee James Thorne

KAIST AI

{jiwoo_hong, noah.lee, thorne}@kaist.ac.kr

SimPO: Simple Preference Optimization with a Reference-Free Reward

Yu Meng[1*] Mengzhou Xia[2*] Danqi Chen[2]

[1]Computer Science Department, University of Virginia

[2]Princeton Language and Intelligence (PLI), Princeton University

RAINBOW PO: A UNIFIED FRAMEWORK FOR COMBINING IMPROVEMENTS IN PREFERENCE OPTIMIZATION

Hanyang Zhao[1*], Genta Indra Winata[2*], Anirban Das[2*], Shi-Xiong Zhang[2], David D. Yao[1], Wenpin Tang[1], Sambit Sahu[2]

[1]Columbia University [2]Capital One

Formalize as discrete reasoning problems

The main technical problem we study

Loss Function

$$-\log \sigma\left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right)$$

The main technical problem we study

Loss Function

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

The main technical problem we study

Loss Function

$$-\log \sigma\left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

- **Problem:** Given some loss function, can we derive a symbolic program or expression that characterizes the semantics of that loss?

The main technical problem we study

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

Loss Function

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

High-level model behavior

- **Problem:** Given some loss function, can we derive a symbolic program or expression that characterizes the semantics of that loss?

The main technical problem we study

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

Loss Function

$$-\log \sigma\left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

Decompilation ⟷ Compilation

- **Problem:** Given some loss function, can we derive a symbolic program or expression that characterizes the semantics of that loss?

The main technical problem we study

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

Loss Function

$$-\log \sigma\left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

Decompilation ⟷ Compilation

- **Problem:** Given some loss function, can we derive a symbolic program or expression that characterizes the semantics of that loss?

The main technical problem we study

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

Loss Function

$$-\log \sigma\left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right)$$

Decompilation ⟷ Compilation

- **Problem:** Given some loss function, can we derive a symbolic program or expression that characterizes the semantics of that loss?

  1. **Compilation**: Translating specifications into loss, well studied.

The main technical problem we study

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

Loss Function

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

Decompilation ⟷ Compilation

- **Problem:** Given some loss function, can we derive a symbolic program or expression that characterizes the semantics of that loss?
  1. **Compilation**: Translating specifications into loss, well studied.
  2. **Decompilation**:Losses to specifications (inverse), less explored.

The main technical problem we study

Symbolic Program | **Loss Function**

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

$\longleftrightarrow$

$$-\log \sigma\left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right)$$

How does this work? Neuro-symbolic techniques

Decompilation $\longleftarrow\longrightarrow$ Compilation

- **Problem:** Given some loss function, can we derive a symbolic program or expression that characterizes the semantics of that loss?
  1. **Compilation**: Translating specifications into loss, well studied.
  2. **Decompilation**:Losses to specifications (inverse), less explored.

The main technical problem we study

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

Loss Function

$$-\log \sigma\left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right)$$

Decompilation ⟷ Compilation

Why is this useful to do?

- **Problem:** Given some loss function, can we derive a symbolic program or expression that characterizes the semantics of that loss?
  1. **Compilation**: Translating specifications into loss, well studied.
  2. **Decompilation**:Losses to specifications (inverse), less explored.

Deriving new losses symbolically, from first principles

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

DPO Loss

$$-\log \sigma\left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right)$$

Deriving new losses symbolically, from first principles

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

modify

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  M(x,y_w)
)
```

DPO Loss

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

Deriving new losses symbolically, from first principles

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

DPO Loss

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

modify

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  M(x,y_w)
)
```

Novel loss

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)(1-\pi_\theta(y_l|x))}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)(1-\pi_\theta(y_w|x))}\right)$$

Deriving new losses symbolically, from first principles

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

DPO Loss

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

modify

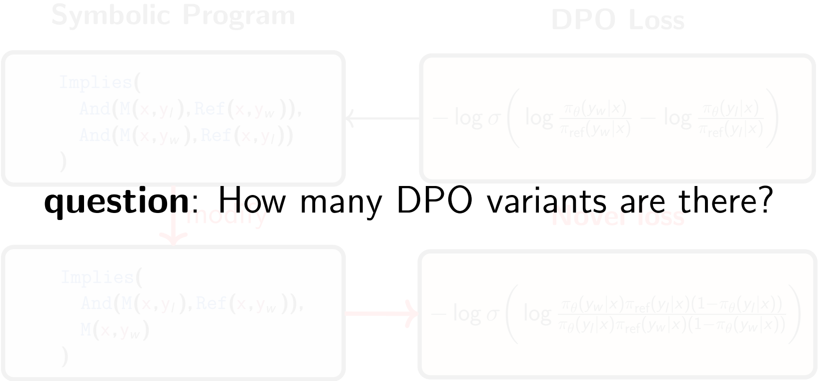
```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  M(x,y_w)
)
```

Novel loss

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)(1-\pi_\theta(y_l|x))}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)(1-\pi_\theta(y_w|x))}\right)$$

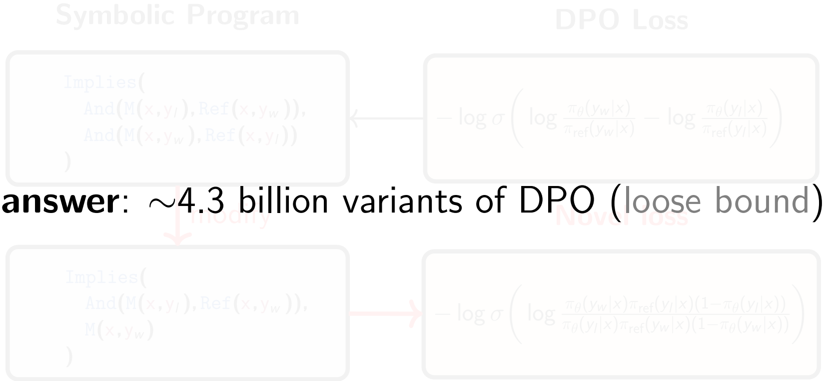
- **Basic questions**: Allows us to better understand the size and structure of the target loss space.

Deriving new losses symbolically, from first principles



question: How many DPO variants are there?

Deriving new losses symbolically, from first principles



answer: $\sim$4.3 billion variants of DPO (loose bound)

Deriving new losses symbolically, from first principles

Symbolic Program

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  And(M(x,y_w),Ref(x,y_l))
)
```

DPO Loss

$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

question: How is this space structured?

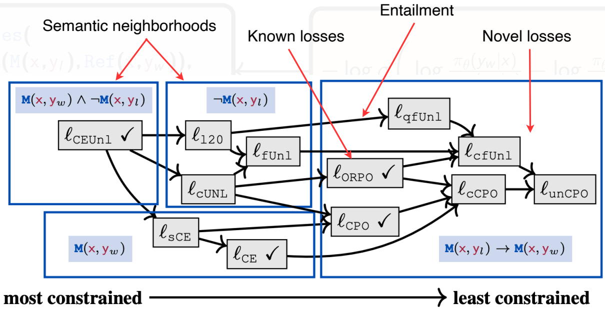
modify

Novel loss

```
Implies(
  And(M(x,y_l),Ref(x,y_w)),
  M(x,y_w)
)
```

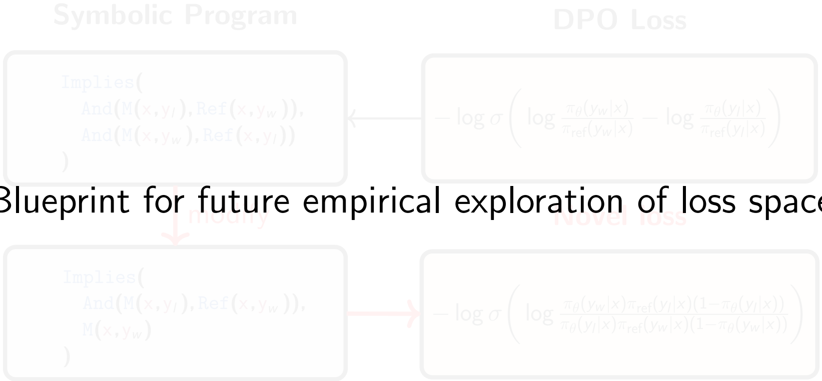
$$-\log\sigma\left(\log\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)(1-\pi_\theta(y_l|x))}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)(1-\pi_\theta(y_w|x))}\right)$$

Deriving new losses symbolically, from first principles



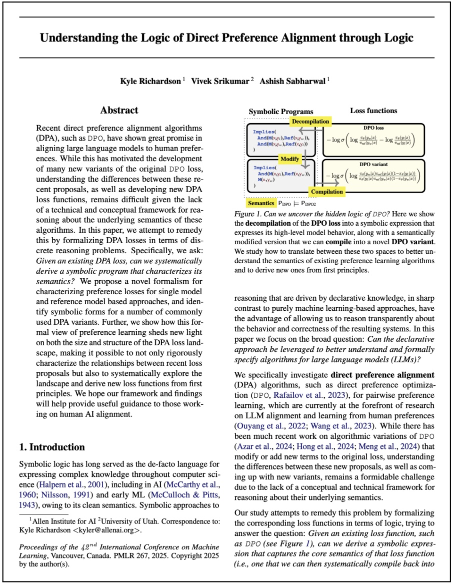
Loss lattice, semantic structure of space, ordering.

Deriving new losses symbolically, from first principles



Blueprint for future empirical exploration of loss space

Please see paper and poster for more details



Understanding the Logic of Direct Preference Alignment through Logic

Kyle Richardson [1] Vivek Srikumar [2] Ashish Sabharwal [1]

Abstract

Recent direct preference alignment algorithms (DPA), such as DPO, have shown great promise in aligning large language models to human preferences. While this has motivated the development of many new variants of the original DPO loss, understanding the differences between these recent proposals, as well as developing new DPA loss functions, remains difficult given the lack of a technical and conceptual framework for reasoning about the underlying semantics of these algorithms. In this paper, we attempt to remedy this by formalizing DPA losses in terms of discrete reasoning problems. Specifically, we ask: *Given an existing DPA loss, can we systematically derive a symbolic program that characterizes its semantics?* We propose a novel formalism for characterizing preference losses for single model and reference model based approaches, and identify symbolic forms for a number of commonly used DPA variants. Further, we show how this formal view of preference learning sheds new light on both the size and structure of the DPA loss landscape, making it possible to not only rigorously characterize the relationships between recent loss proposals but also to systematically explore the landscape and derive new loss functions from first principles. We hope our framework and findings will help provide useful guidance to those working on human AI alignment.

1. Introduction

Symbolic logic has long served as the de-facto language for expressing complex knowledge throughout computer science (Halpern et al., 2001), including in AI (McCarthy et al., 1960; Nilsson, 1991) and early ML (McCulloch & Pitts, 1943), owing to its clean semantics. Symbolic approaches to

Figure 1. Can we uncover the hidden logic of DPO*?* Here we show the **decompilation** of the **DPO loss** into a symbolic expression that expresses its high-level model behavior, along with a semantically modified version that we can **compile** into a novel **DPO variant**. We study how to translate between these two spaces to better understand the semantics of existing preference learning algorithms and to derive new ones from first principles.

reasoning that are driven by declarative knowledge, in sharp contrast to purely machine learning-based approaches, have the advantage of allowing us to reason transparently about the behavior and correctness of the resulting systems. In this paper we focus on the broad question: *Can the declarative approach be leveraged to better understand and formally specify algorithms for large language models (LLMs)?*

We specifically investigate **direct preference alignment** (DPA) algorithms, such as direct preference optimization (DPO, Rafailov et al., 2023), for pairwise preference learning, which are currently at the forefront of research on LLM alignment and learning from human preferences (Ouyang et al., 2022; Wang et al., 2023). While there has been much recent work on algorithmic variations of DPO (Azar et al., 2024; Hong et al., 2024; Meng et al., 2024) that modify or add new terms to the original loss, understanding the differences between these new proposals, as well as coming up with new variants, remains a formidable challenge due to the lack of a conceptual and technical framework for reasoning about their underlying semantics.

Our study attempts to remedy this problem by formalizing the corresponding loss functions in terms of logic, trying to answer the question: *Given an existing loss function, such as* DPO *(see Figure 1), can we derive a symbolic expression that captures the core semantics of that loss function (i.e., one that we can then systematically compile back into*

[1]Allen Institute for AI [2]University of Utah. Correspondence to: Kyle Richardson <kyler@allenai.org>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Extended paper

More detailed slides

References I

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. (2024). Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., and Yang, Y. (2024). Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.