Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization

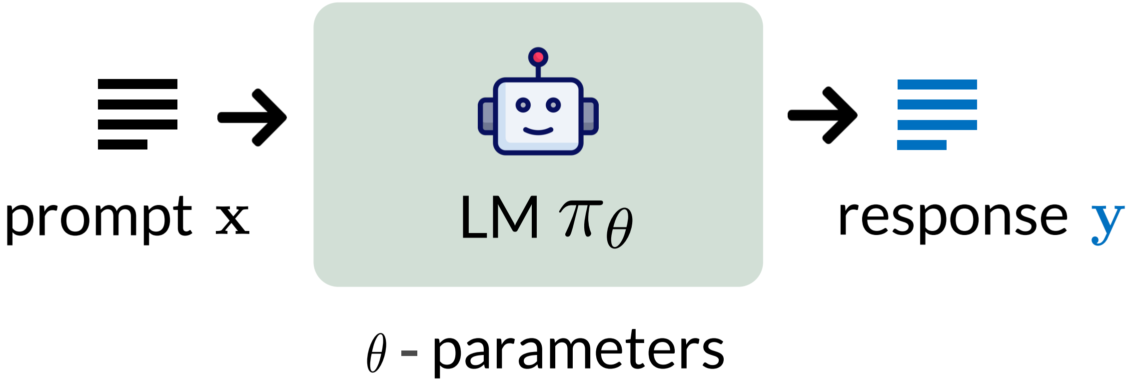
Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, Boris Hanin

Princeton Language and Intelligence, Princeton University

ICLR 2025

Language Models

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



Finetuning LMs via Preference Data

To ensure LMs generate safe and helpful content, they are aligned with human preferences

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 $\mathbf{y}^-$

Main Approaches:

- Reinforcement Learning (e.g. Ouyang et al. 2022)
- Direct Preference Learning (e.g. Rafailov et al. 2023)

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)
- Expensive in terms of memory and compute

Direct Preference Learning

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

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

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

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

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

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

We empirically demonstrate that likelihood displacement can be catastrophic and cause **unintentional unlignment**

Theory: Likelihood displacement is driven by preferences that induce similar embeddings

Based on our theory, we propose a preference similarity measure that allows mitigating likelihood displacement through data filtering

Our work highlights the importance of curating data with distinct preferences, for which our similarity measure may prove valuable