Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization

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

Princeton Language & Intelligence
Princeton ORFE

Direct Preference Learning

(e.g. DPO; Rafailov et al. 2023)

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

↓

Language Model π_θ

↓

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

Likelihood Displacement

Definition: When the probability of $\mathbf{y}^+$ decreases during training

Benign

$\mathbf{z}$ is as preferable as $\mathbf{y}^+$

Catastrophic

$\mathbf{z}$ is substantially less preferable than $\mathbf{y}^+$

Likelihood displacement is prevalent, yet not well understood

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

Question #1

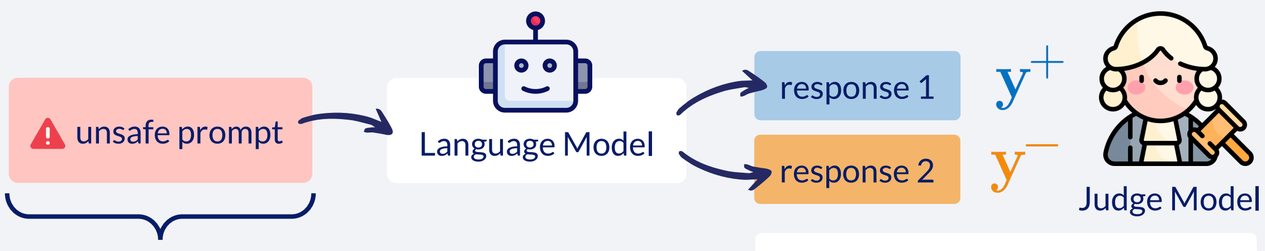
Why does likelihood displacement occur?

Question #2

What are its implications?

1 Likelihood Displacement Can Cause Unintentional Unalignment

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



Results

Probability shifts from preferred refusals to harmful responses!

E.g., the refusal rate of Llama-3-8B-Instruct drops from 74.4% to 33.4%

2 Theory: Likelihood Displacement is Driven by the Embedding Geometry

Approach: Characterize evolution of log probabilities

Main Takeaway

Preferences with similar hidden embeddings lead to likelihood displacement!

Similarity measured by the Centered Hidden Embedding Similarity (CHES) score

CHES score of $(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-)$

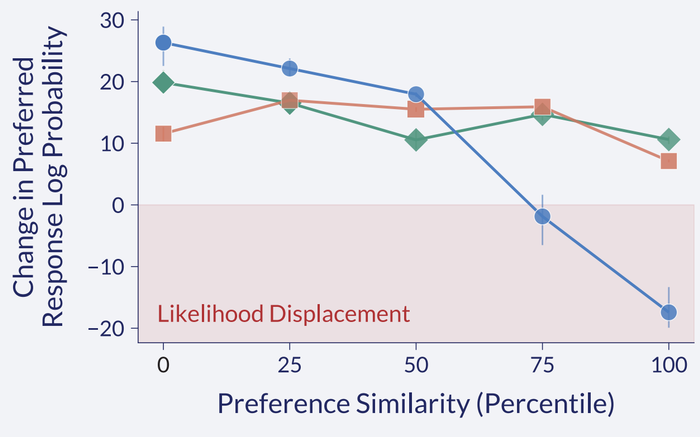
$$\left\langle \sum_{k=1}^{|\mathbf{y}^+|} \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k}} , \sum_{k'=1}^{|\mathbf{y}^-|} \mathbf{h}_{\mathbf{x},\mathbf{y}^-_{<k'}} \right\rangle - \left\| \sum_{k=1}^{|\mathbf{y}^+|} \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<k}} \right\|^2$$

$\mathbf{y}^+$ hidden embeddings — $\mathbf{y}^-$ hidden embeddings

3 Identifying Sources of Likelihood Displacement

Main Takeaway

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



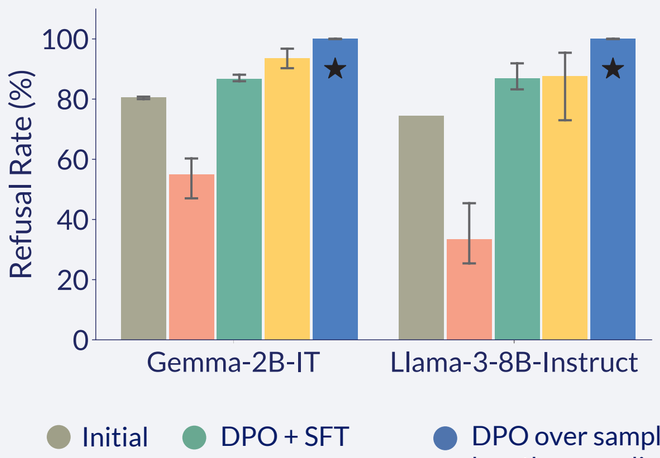
Setting: Llama-3-8B trained via DPO on UltraFeedback subsets

Paper includes similar results for the OLMo-1B and Gemma-2B models and AlpacaFarm dataset

4 Data Filtering via CHES Score

Main Takeaway

Removing samples with high CHES scores mitigates unintentional unalignment



Setting: The same as in 1 — train a language model to refuse unsafe prompts via DPO