

Direct Preference Optimization (DPO)

What is DPO?

DPO directly optimizes the language model using preference data, eliminating the need for a separate reward model and RL training loop.

DPO vs PPO: Architecture Comparison


PPO


Two-Stage Process

Stage 1: Train Reward Model

 Preference Data → $r(\text{output})$

Stage 2: RL Optimization

 Actor-Critic + KL penalty

 Complex hyperparameters


VS


DPO


Single-Stage Process

Stage 1: Direct Optimization

✓ Preference Data → Policy Update

 No separate reward model

 Simpler training pipeline

 Lower memory requirements