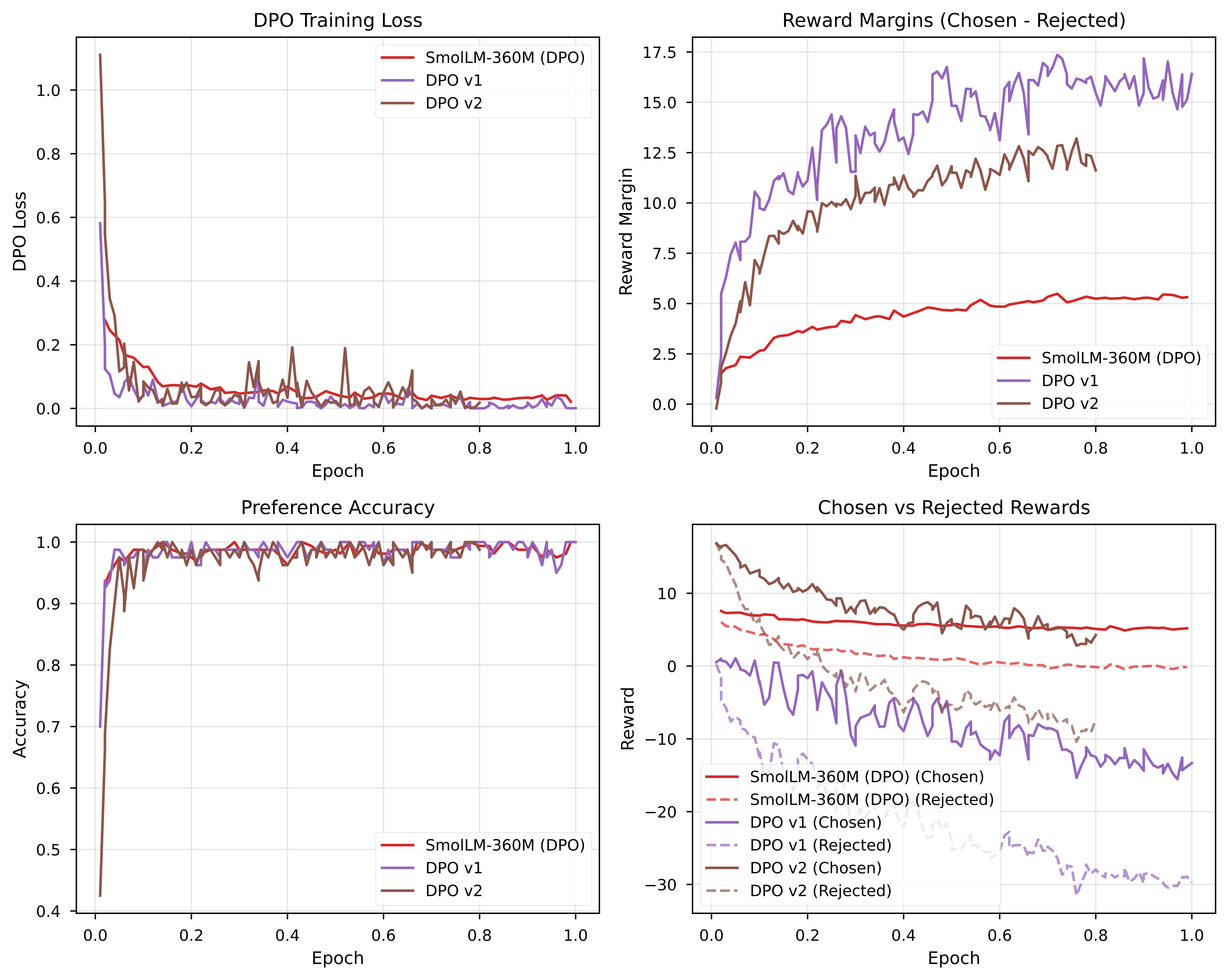
DPO Training Loss
DPO Loss
Epoch
SmolLM-360M (DPO)
DPO v1
DPO v2
Reward Margins (Chosen - Rejected)
Reward Margin
Epoch
Preference Accuracy
Accuracy
Chosen vs Rejected Rewards
Reward
SmolLM-360M (DPO) (Chosen)
SmolLM-360M (DPO) (Rejected)
DPO v1 (Chosen)
DPO v1 (Rejected)
DPO v2 (Chosen)
DPO v2 (Rejected)