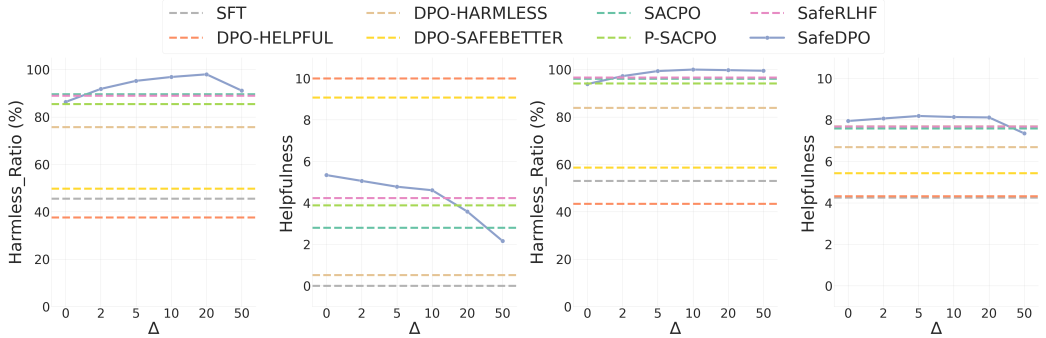

SFT
DPO-HELPFUL
DPO-HARMLESS
DPO-SAFEBETTER
SACPO
P-SACPO
SafeRLHF
SafeDPO
Harmless_Ratio (%)
Helpfulness
Δ
(a) Model-Based Evaluation
(b) GPT-4 Evaluation