Transfer From Ensemble of VLMs to New VLM

Cross Entropy Loss

Cross Entropy Loss vs Gradient Step

Claude 3 Opus

Harmful-Yet-Helpful vs Gradient Step
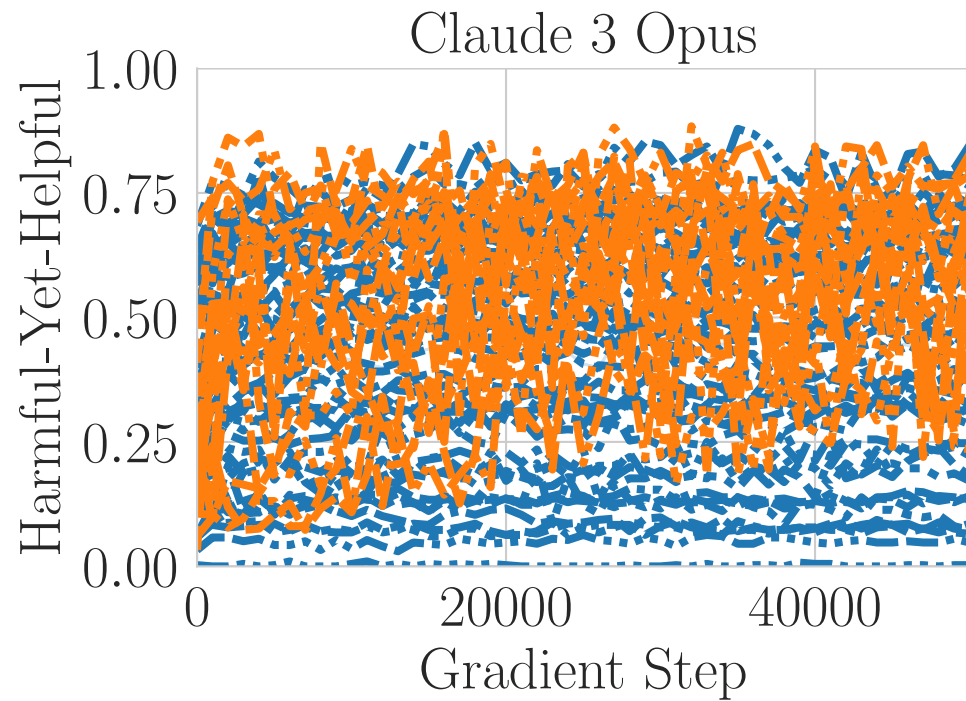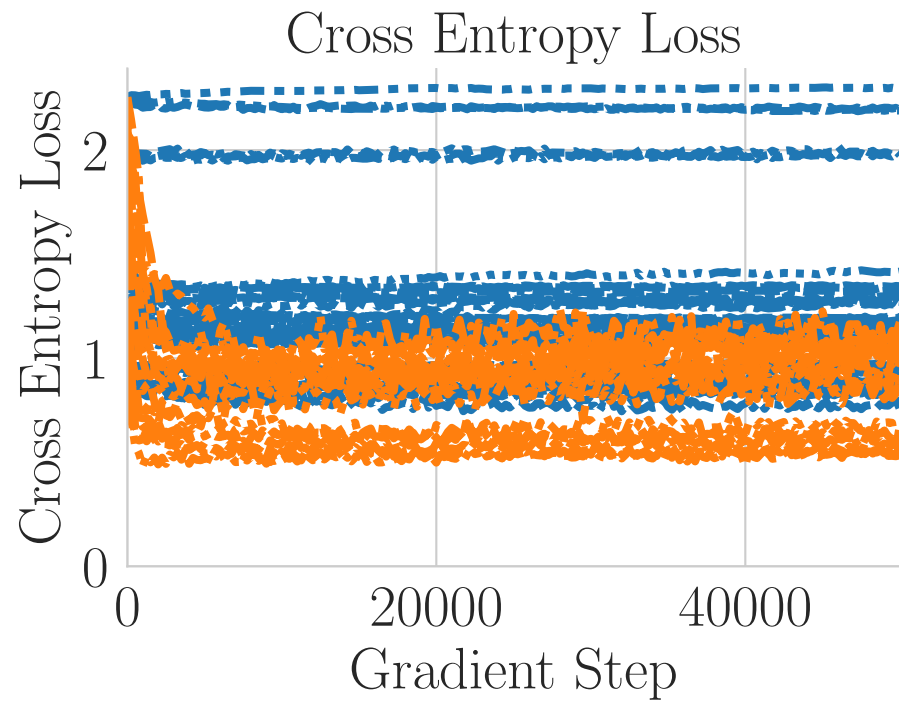
Eval VLM in
Attacked Ensemble

— False

— True

Single VLM Eval Run