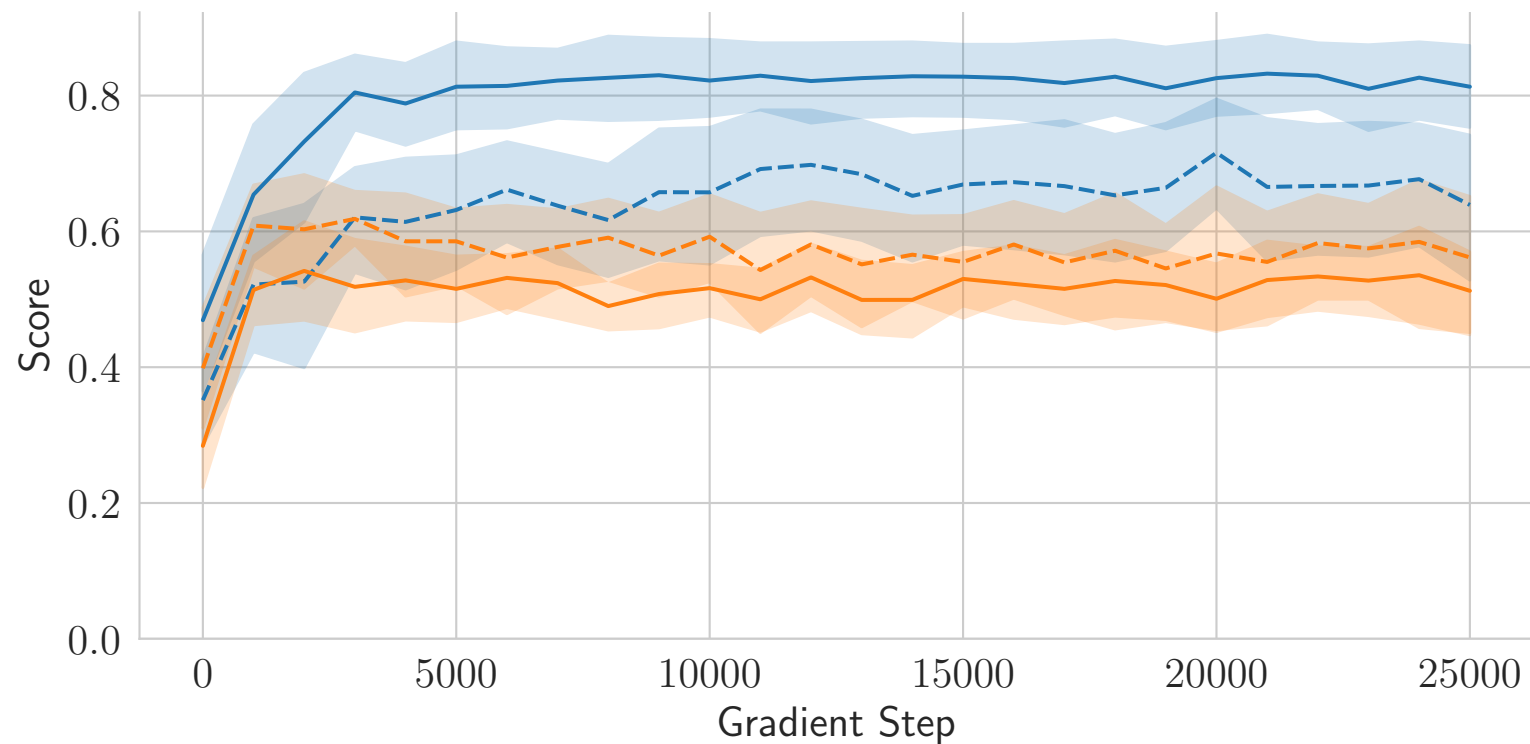


# Universality of Image Jailbreaks

LlamaGuard2 Score



- Attack Dataset (Train Split)
- advbench
  - rylan\_anthropic\_hhh
- Same Data Distribution
- True
  - False