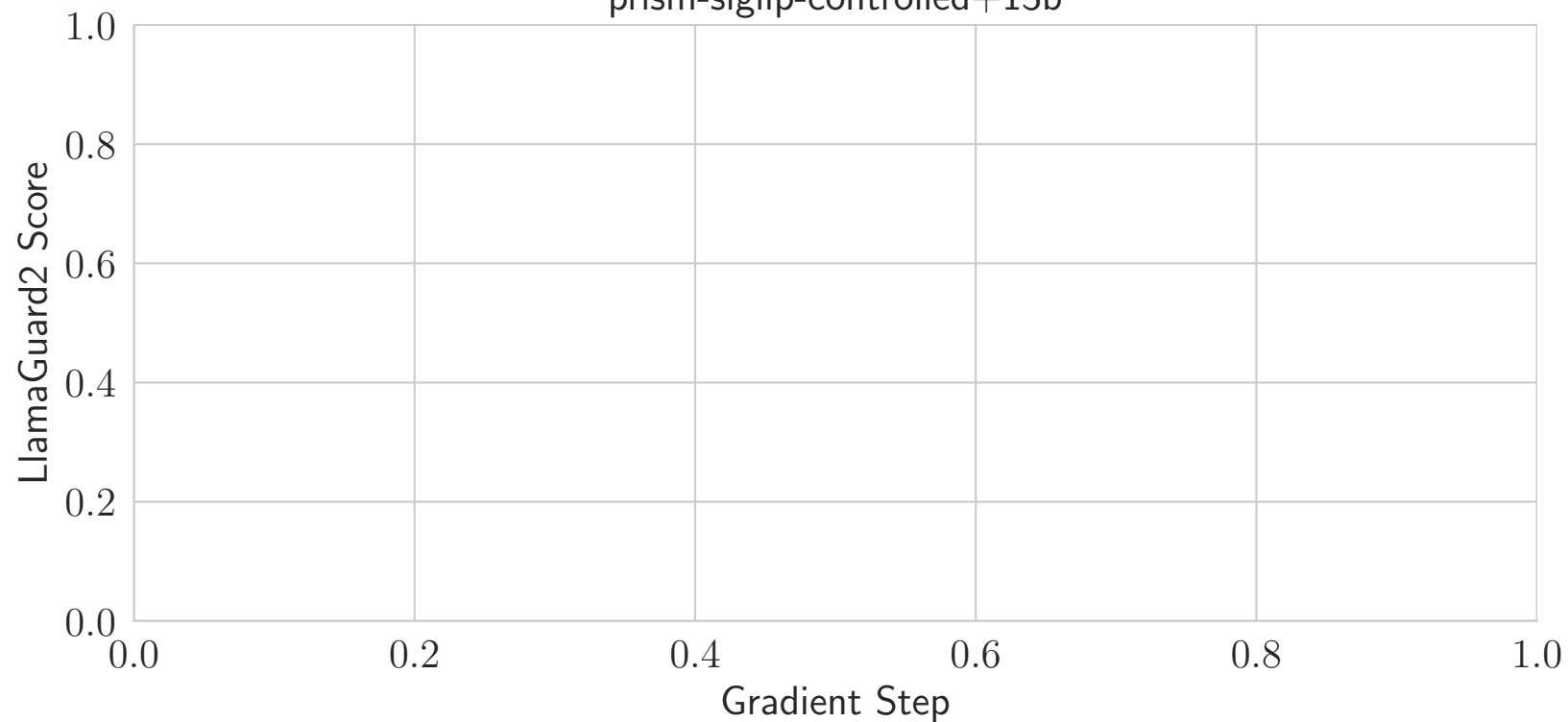


# Attacked Model(s)

'prism-siglip-controlled+13b'



## Evaluated Model

- 'prism-clip+13b'
  - 'prism-clip+7b'
  - 'prism-clip-controlled+7b'
  - 'prism-dinosiglip+7b'
  - 'prism-dinosiglip-controlled+7b'
  - 'prism-reproduction-llava-v15+13b'
  - 'prism-reproduction-llava-v15+7b'
  - 'prism-siglip+7b'
  - 'prism-siglip-controlled+13b'
  - 'prism-siglip-controlled+7b'
- Attack Dataset
- advbench
  - rylan\_anthropic\_hhh