Attacked Ensemble of N = 8 VLMs

-0.8

i. 9. 9. Harmful-Yet-Helpful

-0.2

-0.0