Alternating Classification Non-alternating Unablated 0.5 0.0 -0.5 r = 0.50-1.0 DO preference -1.5 **GPT-2 Small** 0 r = 0.59-5 0.6 0.4

0.2

Human Judgment