

Adversarial Type	Adversarial Input Similarity	Generated Response Similarity
Politics		
Character Edit	$(.94 \pm .04 ; 2.99 \pm .74 ; 4.56 \pm .19)$	$(0.75 \pm .17 ; 2.38 \pm 1.42 ; 3.30 \pm 1.01)$
Paraphrased	$(.85 \pm .13 ; 2.35 \pm .65 ; 3.86 \pm 0.84)$	$(.77 \pm .18 ; 2.52 \pm 1.51 ; 3.39 \pm .99)$
Movies		
Character Edit	$(.93 \pm .05 ; 2.88 \pm .66 ; 4.52 \pm 0.23)$	$(.67 \pm .19 ; 2.06 \pm 1.11 ; 2.75 \pm 1.07)$
Paraphrased	$(.84 \pm .12 ; 2.68 \pm .69 ; 3.97 \pm .76)$	$(0.64 \pm .17 ; 1.87 \pm .76 ; 2.61 \pm .94)$