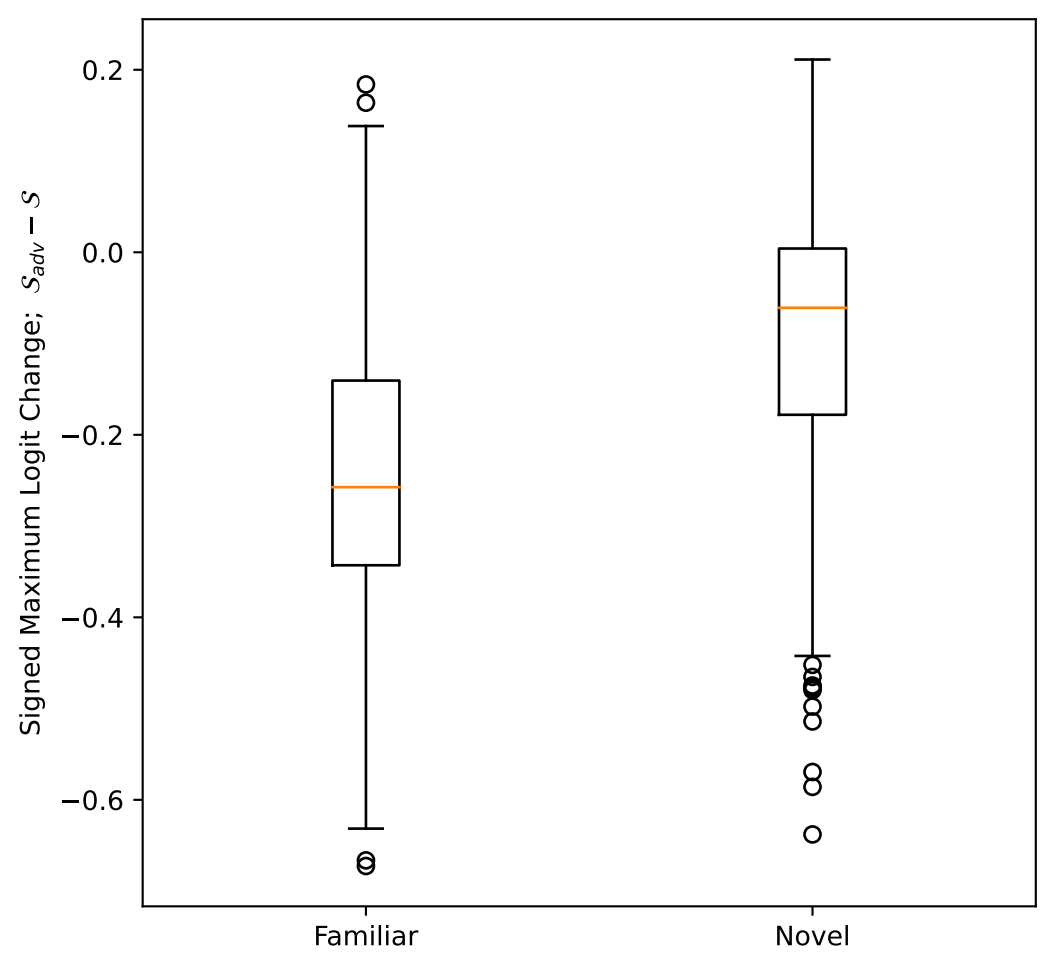
0.2
0.0
−0.2
−0.4
−0.6
Signed Maximum Logit Change; $\mathcal{S}_{adv} - \mathcal{S}$
Familiar
Novel