

Signed Maximum Logit Change; $S_{adv} - S$

0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2

Familiar

Novel

