Pairwise Comparison of Systems (Mean Preference Score) Compared to System A seed 0.6 jailbreak 0.5 0.4 Mean Preference Score 0.3 0.2 0.1

