XGB Model trained with samples created by text-davinci-003 rating answers created by text-davinci-003 annotated by gpt-3.5-turbo used for testing 0.6 -0.5 0.4 % O.3 0 0.2 0.1 0.0 0 50 100 400 800 1600 3200 200 Number of samples

Each box represents exactly 6 kappa values