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 0.25 0.20 0.15 QWK 0.10 0.05 0 0.00 50 100 200 400 800 1600 3200 Number of samples

Each box represents exactly 6 kappa values