BERT Model trained with samples created by human experts rating answers created by text-davinci-003 annotated by gpt-3.5-turbo used for testing 0.08 0.06 0.04 -QWK 0.02 0 0.00 0 -0.02-0.04 -100 200 400 800 1600 50 Number of samples

Each box represents exactly 6 kappa values