BERT Model trained with samples created by human experts rating answers created by gpt4 used for testing 0 0.3 0.2 QWK 0.1 0.0 0 -0.10 0 50 100 200 400 800 1600 Number of samples

Each box represents exactly 6 kappa values