BERT Model trained with samples created by text-davinci-003 rating answers created by human experts used for testing 0.6 0.5 0.4 % 0.3 0 0.2 0.1 0.0 50 100 400 1600 200 800 3200 Number of samples

Each box represents exactly 6 kappa values