BERT Model trained with samples created by gpt4 rating answers created by gpt4 used for training 0.9 0.8 0.7 0.6 ₩ 0.5 0.4 0.3 0.2 50 100 400 800 1600 3200 200 Number of samples Each box represents exactly 6 kappa values