XGB Model trained with samples created by gpt4 rating answers created by gpt4 used for training 8.0 0.6 0.4 0.2 100 400 800 1600 3200 50 200 Number of samples Each box represents exactly 6 kappa values