XGB Model trained with samples created by gpt4 rating answers created by gpt4 used for testing 0.40 -0.35 -0.30 -0.25 -¥ 0.20 − 0.15 0.10 0.05 -0.00 -100 800 1600 3200 50 200 400 Number of samples Each box represents exactly 6 kappa values