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