XGB Model trained with samples created by human experts rating answers created by gpt4 used for testing 0.150 -0.125 -0.100 Kappa 0.075 0.050 -0.025 0.000 -0.025300 400 500 100 200 600 700 800 Amount of samples the model was trained with Each value represents the agreement of exactly 1600 ratings