XGB Model trained with samples created by gpt4 rating answers created by text-davinci-003 annotated by gpt-3.5-turbo used for testing 0.6 -0.5 -0.4 Kappa 0.3 0.2 -0.1 300 400 500 600 700 800 100 200 Amount of samples the model was trained with Each value represents the agreement of exactly 1576 ratings