XGB Model trained with samples created by text-davinci-003 annotated by gpt-3.5-turbo rating answers created by human experts used for testing 0.6 -0.5 0.4 -Kappa c.0 0.2 -0.1 1000 1500 3000 500 2000 2500 Amount of samples the model was trained with Each value represents the agreement of exactly 392 ratings