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 0.50 -0.45 Kappa - 08.0 0.35 0.30 0.25 -1000 3000 500 1500 2000 2500 Amount of samples the model was trained with Each value represents the agreement of exactly 392 ratings