XGB Model trained with samples created by text-davinci-003 annotated by gpt-3.5-turbo rating answers created by text-davinci-003 annotated by gpt-3.5-turbo used for testing 0.16 -0.14 0.12 Kappa 0.10 0.08 -0.06 -0.04 0.02 700 800 100 200 300 400 500 600 Amount of samples the model was trained with Each value represents the agreement of exactly 1576 ratings