BERT Model trained with samples created by gpt4 rating answers created by text-davinci-003 used for testing 0.3 0.2 Kappa 0.1 0.0 -0.11500 2000 500 1000 2500 3000 Amount of samples the model was trained with Each value represents the agreement of exactly 1577 ratings