BERT 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 0.14 -0.12 -0.10 -Kappa - 80.0 0.06 -0.04 -0.02 -0.00

400

Amount of samples the model was trained with

500

100

200

300

Each value represents the agreement of exactly 1576 ratings

600

700

800